

Subspace Perspective on Canonical Correlation Analysis: Dimension Reduction and Minimax Rates

Zhuang Ma* Xiaodong Li†

May 13, 2016

Abstract

Canonical correlation analysis (CCA) is a fundamental statistical tool for exploring the correlation structure between two sets of random variables. In this paper, motivated by recent success of applying CCA to learn low dimensional representations of high dimensional objects, we propose to quantify the estimation loss of CCA by the excess prediction loss defined through a prediction-after-dimension-reduction framework. Such framework suggests viewing CCA estimation as estimating the subspaces spanned by the canonical variates. Interestingly, the proposed error metrics derived from the excess prediction loss turn out to be closely related to the principal angles between the subspaces spanned by the population and sample canonical variates respectively.

We characterize the non-asymptotic minimax rates under the proposed metrics, especially the dependency of the minimax rates on the key quantities including the dimensions, the condition number of the covariance matrices, the canonical correlations and the eigen-gap, with minimal assumptions on the joint covariance matrix. To the best of our knowledge, this is the first finite sample result that captures the effect of the canonical correlations on the minimax rates.

1 Introduction

Canonical correlation analysis (CCA), first introduced by [Hotelling \[1936\]](#), is a fundamental statistical tool to characterize the relationship between two groups of random variables and finds a wide range of applications across many different fields. For example, in genome-wide association study (GWAS), CCA is used to discover the genetic associations between the genotype data of single nucleotide polymorphisms (SNPs) and the phenotype data of gene expression levels [[Witten et al., 2009](#); [Chen et al., 2012](#)]. In information retrieval, CCA is used to embed both the search space (e.g. images) and the query space (e.g. text) into a shared low dimensional latent space such that the similarity between the queries and the candidates can be quantified [[Rasiwasia et al., 2010](#); [Gong et al., 2014](#)]. In natural language processing, CCA is applied to the word co-occurrence matrix, which learns a vector representation of the words that is able to capture the semantics [[Dhillon et al.,](#)

*Z. Ma is with Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA. zhuangma@wharton.upenn.edu.

†X. Li is with the Statistics Department at the University of California, Davis, CA. xdgli@ucdavis.edu.

2011; Faruqui and Dyer, 2014]. Other applications, to name a few, include fMRI data analysis [Friman et al., 2003], computer vision [Kim et al., 2007] and speech recognition [Arora and Livescu, 2013; Wang et al., 2015].

The enormous empirical success motivates us to revisit the estimation problem of canonical correlation analysis. From a decision-theoretic point of view, two questions are naturally posed: What is the proper error metric to quantify the discrepancy between population CCA and its sample estimates? And under such metric, what are the quantities that characterize the fundamental statistical limits?

The justification of loss functions, in the context of CCA, has seldom appeared in the literature. Start from the first principle that the proper metric to quantify the estimation loss should depend on the specific purpose of using CCA, we find that the applications discussed at the beginning mainly fall into two categories: identify variables of interest and dimension reduction. The first category, mostly in genomic research [Witten et al., 2009; Chen et al., 2012], treats one group of variables as responses and the other group as covariates. The goal is to discover the specific subset of the covariates that are most correlated with the responses. Such applications are featured by low signal-to-noise ratio and the interpretability of the results is the major concern. The other category is investigated extensively in statistical machine learning and engineering community where CCA is used to learn low dimensional latent representations of complex objects such as images [Rasiwasia et al., 2010], text [Dhillon et al., 2011] and speeches [Arora and Livescu, 2013]. These scenarios are usually accompanied with relatively high signal-to-noise ratio and the prediction accuracy, using the learned low dimensional embeddings as the new set of predictors, is of primary interest. In recent years, there has been a series of publications establishing fundamental theoretical guarantees for CCA to achieve sufficient dimension reduction [Kakade and Foster, 2007; Foster et al., 2008; Sridharan and Kakade, 2008; Fukumizu et al., 2009; Chaudhuri et al., 2009], especially under a multi-view setup as will be discussed more detailedly in Section 2.4.

In this paper, we aim to address the problems raised above by treating CCA as a tool for dimension reduction.

1.1 Canonical Correlation Analysis and Linear Invariance

On the population level, CCA is designed to extract the most correlated directions between two sets of random variables: $\mathbf{x} \in \mathbb{R}^{p_1}$ and $\mathbf{y} \in \mathbb{R}^{p_2}$. Specifically, CCA recursively finds the pairs of vectors $\phi_i \in \mathbb{R}^{p_1}, \psi_i \in \mathbb{R}^{p_2}, 1 \leq i \leq p := \min\{p_1, p_2\}$ such that

$$\begin{aligned} (\phi_i, \psi_i) &= \arg \max_{\phi^\top \Sigma_x \phi = 1, \psi^\top \Sigma_y \psi = 1} \phi^\top \Sigma_{xy} \psi \\ \text{subject to } &\phi^\top \Sigma_x \phi_j = 0, \psi^\top \Sigma_y \psi_j = 0, \forall 1 \leq j \leq i-1. \end{aligned} \quad (1.1)$$

For $1 \leq i \leq p$, (ϕ_i, ψ_i) are the canonical coefficients (or loadings), $(\phi_i^\top \mathbf{x}, \psi_i^\top \mathbf{y})$ are the canonical variates and $\lambda_i := \phi_i^\top \Sigma_x \psi_i$ are the canonical correlations. Let $\Phi = [\phi_1, \dots, \phi_p]$ and $\Psi = [\psi_1, \dots, \psi_p]$ be the loading matrices. With these notations, the first type of applications discussed above can be understood as identifying the support of the top- k canonical vectors: $\Phi_{1:k}$ and $\Psi_{1:k}$, where $\Phi_{1:k} \in \mathbb{R}^{p_1 \times k}$ and $\Psi_{1:k} \in \mathbb{R}^{p_2 \times k}$ consist of the first k columns of Φ and Ψ respectively. While for dimension reduction, which motivates the

paper, is concerned with the leading k canonical variates: $\Phi_{1:k}^\top \mathbf{x}$ and $\Psi_{1:k}^\top \mathbf{y}$ (k is assumed to be known *a priori*, or a pre-specified budget).

What distinguishes CCA from other dimension reduction methods like principal component analysis or partial least squares is its linear invariance. As highlighted in Hotelling [1936] when canonical correlation analysis was first developed:

“The relations between two sets of variates with which we shall be concerned are those that remain invariant under internal linear transformations of each sets separately.”

Among all the population parameters, Hotelling [1936] noticed that the canonical correlations $\lambda_1, \dots, \lambda_p$ and the functions of these quantities are the only linear invariants of the system. On the contrary, the loadings Φ and Ψ will change accordingly either with rotation of axes or scaling of the variables, which diminishes the rationale for using an error metric built directly upon the loadings. If extending Hotelling’s notion of invariants to include random vectors, the canonical variates are actually invariant under linear transformations of each sets separately. To illustrate, let $\mathbf{T}_1, \mathbf{T}_2$ be any two nonsingular matrices and define the new random vectors $\mathbf{a} = \mathbf{T}_1^\top \mathbf{x}, \mathbf{b} = \mathbf{T}_2^\top \mathbf{y}$. As will be shown in Section 3.1, $\mathbf{T}_1^{-1} \Phi_{1:k}, \mathbf{T}_2^{-1} \Psi_{1:k}$ are the top- k canonical coefficients of (\mathbf{a}, \mathbf{b}) . Therefore, the top- k canonical variates of (\mathbf{a}, \mathbf{b}) will be $(\mathbf{T}_1^{-1} \Phi_{1:k})^\top \mathbf{a} = \Phi_{1:k}^\top \mathbf{x}$ and $(\mathbf{T}_2^{-1} \Psi_{1:k})^\top \mathbf{b} = \Psi_{1:k}^\top \mathbf{y}$, which are the same as those of (\mathbf{x}, \mathbf{y}) . This fact substantiates our interest in the canonical variates instead of the loadings. Let $(\hat{\Phi}_{1:k}, \hat{\Psi}_{1:k})$ be any generic estimator of the loadings. Then the two questions that we aim to answer can be recast as:

What is the proper error metric to quantify the discrepancy between $(\Phi_{1:k}^\top \mathbf{x}, \Psi_{1:k}^\top \mathbf{y})$ and its sample estimates $(\hat{\Phi}_{1:k}^\top \mathbf{x}, \hat{\Psi}_{1:k}^\top \mathbf{y})$? And under such metric, what are the quantities that characterize the fundamental statistical limits?

For the rest of the paper, we will only focus on the relationship between $\hat{\Phi}_{1:k}^\top \mathbf{x}$ and $\Phi_{1:k}^\top \mathbf{x}$ since similar results can be obtained for the other pair by symmetry.

1.2 Subspace Estimation and Subspace Loss

In Section 2, we show that when CCA is used for dimension reduction, it is the difference between the predictive power of $\Phi_{1:k}^\top \mathbf{x}$ and $\hat{\Phi}_{1:k}^\top \mathbf{x}$ that matters, instead of the distance by simply treating $\Phi_{1:k}^\top \mathbf{x}$ and $\hat{\Phi}_{1:k}^\top \mathbf{x}$ as vectors. We propose to characterize the discrepancy between the predictive power by the excess prediction loss induced by replacing the population canonical variates $\Phi_{1:k}^\top \mathbf{x}$ with the sample estimates $\hat{\Phi}_{1:k}^\top \mathbf{x}$ in terms of predicting a generic response. Specifically, when linear prediction is concerned, such discrepancy is reduced to the difference between the subspaces spanned by the population and sample canonical variates, denoted by $\text{span}(\mathbf{x}^\top \Phi_{1:k})$ and $\text{span}(\mathbf{x}^\top \hat{\Phi}_{1:k})$. This suggests that CCA estimation can be viewed as subspace estimation, that is, estimating the subspace spanned by the leading- k canonical variates: $\text{span}(\mathbf{x}^\top \Phi_{1:k})$. From this perspective, the error metric $\mathcal{L}(\cdot, \cdot)$ we pursue should be rewritten as

$$\mathcal{L}(\Phi_{1:k}^\top \mathbf{x}, \hat{\Phi}_{1:k}^\top \mathbf{x}) := \mathcal{L}(\text{span}(\mathbf{x}^\top \Phi_{1:k}), \text{span}(\mathbf{x}^\top \hat{\Phi}_{1:k})), \quad (1.2)$$

Interestingly, the error metrics derived through the excess prediction loss turn out to be closely related to the principal angles (defined in Section 2.3) between $\text{span}(\hat{\Phi}_{1:k}^\top \mathbf{x})$ and

$\text{span}(\Phi_{1:k}^\top \mathbf{x})$. Suppose $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ is the vector of such principal angles. It is proved in Theorem 2.1 that,

$$\begin{aligned} \text{Worst case excess prediction loss} &\simeq \left\| P_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|^2 = \|\sin(\boldsymbol{\theta})\|_\infty^2 \\ \text{Bayesian excess prediction loss} &\simeq \left\| P_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|_F^2 / 2k = \|\sin(\boldsymbol{\theta})\|_2^2 / k \end{aligned} \quad (1.3)$$

where \simeq means ‘equal up to constant’ and $P_{(\cdot)}$ denotes the projection matrix.

1.3 Minimax Rates

In section 3, we characterize the non-asymptotic minimax estimation rates for CCA under the proposed error metrics in (1.3), especially the dependency of the minimax rates on the key quantities, including the dimensions, the condition number of the covariance matrices, the canonical correlations and the eigen-gap. Informally, with operator norm error as an example, in Theorem 3.2 and Theorem 3.3, we show that under certain sample size condition ($n \geq C_{\lambda_k, \lambda_{k+1}}(p_1 + p_2)$), the minimax rate is characterized by

$$\inf_{\hat{\Phi}_{1:k}} \sup_{\Sigma \in \mathcal{F}} \mathbb{E} \left[\left\| P_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|^2 \right] \asymp \frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2) p_1}{(\lambda_k - \lambda_{k+1})^2 n}.$$

To the best of our knowledge, it is the first finite sample result to capture the factor $(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)$. This term is not negligible because λ_k, λ_{k+1} are parameters depending on the dimensions and should not be treated as constants. In practice, as the number of variables increases, one should expect the canonical correlations to increase as well. The other important feature is the independence of the dimension p_2 . If one is only interested in the ‘estimation’ of the canonical variates of \mathbf{x} , then even in the regime $p_2 \gg p_1$, as long as the sample size is large enough, the minimax rate of ‘estimating’ $\Phi_{1:k}^\top \mathbf{x}$ does not depend on p_2 . This phenomenon was also revealed in Gao et al. [2015] and the recent work of Cai and Zhang [2016] with extra assumption that all the residual canonical correlations are zero: $\lambda_{k+1} = \dots = \lambda_{p_1} = 0$. Finally, the minimax rates are independent of the condition number of the covariance matrices: $\kappa(\Sigma_x), \kappa(\Sigma_y)$. This is due to the linear invariance of the canonical variates as illustrated in Section 3.1. We hope our theoretical findings could provide some guidance for the practical use of CCA because in real applications, all these factors do matter both computationally and statistically [Ma et al., 2015].

Specifically, the upper bound is achieved by sample CCA which is defined in the same manner as (1.1) by replacing the population covariance matrices with the corresponding sample estimates. The sample canonical variates are also linear invariant, which is crucial to reduce the estimation error of sample CCA to the “standard form”, as spelled out in Section 3.1, that the separate covariance matrices are identity and the cross covariance matrix is diagonal.

Theoretical understanding for the estimation of CCA dates back to the study of the asymptotic distribution of sample CCA, in the low dimensional regime with fixed dimensions and sample size going to infinity, for both sample canonical coefficients and sample canonical correlations [Hotelling, 1936; Hsu, 1941; Izenman, 1975; Anderson, 1984, 1999] (and many

others). More recently, [Chen et al. \[2013\]](#) and [Gao et al. \[2014, 2015\]](#) have studied the non-asymptotic minimax rates of sparse CCA in a high dimensional setup. We defer the detailed comparison between these results and ours to Section 3.2.

1.4 Notations

Throughout the paper, we use lower-case and upper-case bold letters to represent vectors and matrices. For any matrix $\mathbf{U} \in \mathbb{R}^{n \times p}$ and vector $\mathbf{u} \in \mathbb{R}^p$, $\|\mathbf{U}\|, \|\mathbf{U}\|_F$ denotes operator norm and Frobenius norm respectively, $\|\mathbf{u}\|$ denotes the vector l_2 norm, $\mathbf{U}_{1:k}$ denotes the submatrix consisting of the first k columns of \mathbf{U} , and $\mathbf{P}_{\mathbf{U}}$ stands for the projection matrix onto the column space of \mathbf{U} . Moreover, we use $\sigma_{\max}(\mathbf{U})$ and $\sigma_{\min}(\mathbf{U})$ to represent the largest and smallest singular value of \mathbf{U} respectively, and $\kappa(\mathbf{U}) = \sigma_{\max}(\mathbf{U})/\sigma_{\min}(\mathbf{U})$ to denote the condition number of the matrix. We use \mathbf{I}_p for the identity matrix of dimension p and $\mathbf{I}_{p,k}$ for the submatrix composed of the first k columns of \mathbf{I}_p . Further, $\mathcal{O}(m, n)$ stands for the set of $m \times n$ matrices with orthonormal columns and \mathbb{S}_+^p denotes the set of $p \times p$ strictly positive definite matrices. For a random vector $\mathbf{x} \in \mathbb{R}^p$, $\text{span}(\mathbf{x}^\top) = \{\mathbf{x}^\top \mathbf{w}, \mathbf{w} \in \mathbb{R}^p\}$ denotes the subspace of all the linear combinations of \mathbf{x} . Other notations will be specified within the corresponding context.

2 Prediction Loss for Dimension Reduction

In this section, we propose two natural loss functions to quantify the discrepancy between the population and sample canonical variates through a prediction-after-dimension-reduction framework.

2.1 Linear Prediction Revisited

First of all, we review the basics of linear model theory under the random design setup. Given

$$x_1, \dots, x_p, z \in L^2(\Omega, \mathcal{F}, \mathcal{P}),$$

where $L^2(\Omega, \mathcal{F}, \mathcal{P})$ is the set of random variables with mean zero and finite second moment. Suppose the goal is to predict the response z with the random vector $\mathbf{x} = (x_1, \dots, x_p)^\top$. We measure the prediction loss by

$$\text{loss}(z|\mathbf{x}) = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathbb{E}[(z - \mathbf{x}^\top \boldsymbol{\beta})^2].$$

We further assume that (\mathbf{x}, z) has joint covariance matrix:

$$\text{Cov} \left(\begin{bmatrix} \mathbf{x} \\ z \end{bmatrix} \right) = \begin{bmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xz} \\ \boldsymbol{\Sigma}_{xz}^\top & \sigma_z^2 \end{bmatrix}.$$

By classical linear model theory

$$\begin{aligned} \boldsymbol{\beta}^* &= \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathbb{E}[(z - \mathbf{x}^\top \boldsymbol{\beta})^2] = \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Sigma}_{xz} \\ \text{loss}(z|\mathbf{x}) &= \sigma_z^2 - \boldsymbol{\Sigma}_{zx} \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Sigma}_{xz} = \sigma_z^2 (1 - \|\mathbf{r}_{xz}\|^2) \end{aligned}$$

where $\mathbf{r}_{xz} = \Sigma_x^{-1} \Sigma_{xz} \sigma_z^{-1}$ is the correlation vector between \mathbf{x} and z , and $\|\mathbf{r}_{xz}\|^2$ is the population R^2 which characterizes the proportion of the variability in the response z explained by the predictor \mathbf{x} . One important feature of such prediction loss is its linear invariance. Define the linear subspace spanned by the coordinates of \mathbf{x} as

$$\text{span}(\mathbf{x}^\top) = \{\mathbf{x}^\top \mathbf{w} : \mathbf{w} \in \mathbb{R}^p\} \subset L^2(\Omega, \mathcal{F}, \mathcal{P}),$$

For another set of random variables $\{v_1, \dots, v_q\} \in L^2(\Omega, \mathcal{F}, \mathcal{P})$, if $\text{span}(\mathbf{v}) = \text{span}(\mathbf{x})$ by denoting $\mathbf{v} = (v_1, \dots, v_q)^\top$, then $\text{loss}(z|\mathbf{x}) = \text{loss}(z|\mathbf{v})$ and this implies that we can rewrite

$$\text{loss}(z|\mathbf{x}) = \text{loss}(z|\text{span}(\mathbf{x}^\top)).$$

These two notations will be used interchangeably throughout the paper. The linear invariance property can be revealed by noticing that $\|\mathbf{r}_{xz}\|^2 = \mathbb{E}[(\mathbf{P}_{\text{span}(\mathbf{x})} z)^2] / \mathbb{E}[z^2]$ where $\mathbf{P}_{(\cdot)}$ is the projection operator defined in the Hilbert space $L^2(\Omega, \mathcal{F}, \mathcal{P})$ with inner product defined as the covariance between two random variables.

2.2 Competing with Oracles

Consider the scenario where the predictor \mathbf{x} is in a high dimensional space and many directions in $\text{span}(\mathbf{x})$ might be redundant in terms of predicting the response z . Practitioners usually perform certain kind of dimension reduction on \mathbf{x} before applying supervised learning algorithms. Suppose $\mathbf{U} \in \mathbb{R}^{p \times k}$ is the reduction matrix obtained by some generic dimension reduction method. The subspace perspective of the prediction loss discussed in previous section suggests $\text{loss}(z|\text{span}(\mathbf{x}^\top \mathbf{U})) - \text{loss}(z|\text{span}(\mathbf{x}))$, or simply $\text{loss}(z|\text{span}(\mathbf{x}^\top \mathbf{U}))$ as the measure of goodness for dimension reduction algorithms.

For any given reduction matrices $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{R}^{p \times k}$, the discrepancy between their prediction loss can be quantified by

$$\text{loss}(z|\text{span}(\mathbf{x}^\top \mathbf{U}_1)) - \text{loss}(z|\text{span}(\mathbf{U}_2^\top \mathbf{x})) = \mathbb{E}[(\mathbf{P}_{\text{span}(\mathbf{x}^\top \mathbf{U}_2)} z)^2 - (\mathbf{P}_{\text{span}(\mathbf{x}^\top \mathbf{U}_1)} z)^2] \quad (2.1)$$

$$= \sigma_z^2 \left(\mathbf{r}_{xz}^\top \left(\mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_2} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_1} \right) \mathbf{r}_{xz} \right) \quad (2.2)$$

as will be proved in Section 5.1. The first equality is geometrically straightforward, measuring the proportion of the variability in response z explained by the two subspaces $\text{span}(\mathbf{U}_1^\top \mathbf{x})$ and $\text{span}(\mathbf{U}_2^\top \mathbf{x})$. The algebraic expression in the second equality is less obvious but decouples the loss into an interaction between a supervised learning factor \mathbf{r}_{xz} and an unsupervised learning factor $\mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_2} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_1}$. To shed more light on this excess risk, we parametrize the joint covariance matrix of (\mathbf{x}, z) in terms of separate covariance matrices Σ_x, σ_z^2 and the correlation vector $\mathbf{r}_{xz} = \text{correlation}(\mathbf{x}, z)$, that is

$$\text{Cov} \left(\begin{bmatrix} \mathbf{x} \\ z \end{bmatrix} \right) = \begin{bmatrix} \Sigma_x & \Sigma_{xz} \\ \Sigma_{xz}^\top & \sigma_z^2 \end{bmatrix} = \begin{bmatrix} \Sigma_x^{1/2} & 0 \\ 0 & \sigma_z \end{bmatrix} \begin{bmatrix} \mathbf{I}_p & \mathbf{r}_{xz} \\ \mathbf{r}_{zx} & 1 \end{bmatrix} \begin{bmatrix} \Sigma_x^{1/2} & 0 \\ 0 & \sigma_z \end{bmatrix}. \quad (2.3)$$

Considering the worst case discrepancy across all possible correlation structures,

$$\sup_{\|\mathbf{r}_{xz}\|^2=R^2} \{ \text{loss}(z|\text{span}(\mathbf{x}^\top \mathbf{U}_1)) - \text{loss}(z|\text{span}(\mathbf{x}^\top \mathbf{U}_2)) \} = \sigma_z^2 R^2 \left\| \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_1} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_2} \right\|, \quad (2.4)$$

which suggests the right hand side of the equation as a sensible metric to quantify the difference between the two reduction matrices. Sometimes, it might be more informative to replace the competitor \mathbf{U}_2 with an oracle reduction matrix, denoted by \mathbf{U}_\star . As suggested by (2.2), we say a reduction matrix \mathbf{U}_\star is an oracle reduction matrix if $\mathbf{P}_{\Sigma_x^{1/2}\mathbf{U}_\star} \mathbf{r}_{xz} = \mathbf{r}_{xz}$. Define $\mathcal{A} = \{\mathbf{r} : \|\mathbf{r}\|^2 = R^2, \mathbf{P}_{\Sigma_x^{1/2}\mathbf{U}_\star} \mathbf{r} = \mathbf{r}\}$ as the set of correlation vectors with fixed population R^2 such that \mathbf{U}_\star is an oracle reduction matrix. If considering the worst case excess prediction loss within \mathcal{A} , as proved in Section 5.1,

$$\sup_{\mathbf{r}_{xz} \in \mathcal{A}} \{loss(z|\text{span}(\mathbf{x}^\top \mathbf{U})) - loss(z|\text{span}(\mathbf{x}^\top \mathbf{U}_\star))\} = \sigma_z^2 R^2 \left\| \mathbf{P}_{\Sigma_x^{1/2}\mathbf{U}} - \mathbf{P}_{\Sigma_x^{1/2}\mathbf{U}_\star} \right\|^2. \quad (2.5)$$

Interestingly, the operator norm is replaced by its square when the competitor is an oracle reduction matrix. On the other hand, from a Bayesian perspective, considering the prior that the correlation vector \mathbf{r}_{xz} is sampled according to the uniform measure (Haar measure) on \mathcal{A} , then it is proved in Section 5.1 that the average excess prediction loss satisfies:

$$\mathbb{E}_{\mathbf{r}_{xz} \sim \pi} \{loss(z|\text{span}(\mathbf{x}^\top \mathbf{U})) - loss(z|\text{span}(\mathbf{x}^\top \mathbf{U}_\star))\} = \frac{\sigma_z^2 R^2}{2k} \left\| \mathbf{P}_{\Sigma_x^{1/2}\mathbf{U}} - \mathbf{P}_{\Sigma_x^{1/2}\mathbf{U}_\star} \right\|_F^2. \quad (2.6)$$

The analysis in this section essentially connects the prediction loss for the response z with the estimation loss for the oracle reduction matrix \mathbf{U}_\star under the metrics derived in (2.5) and (2.6). Therefore, when CCA is used for dimension reduction, it is natural to quantify the discrepancy between $\hat{\Phi}_{1:k}^\top \mathbf{x}$ and $\Phi_{1:k}^\top \mathbf{x}$ by the excess prediction loss:

$$\left\| \mathbf{P}_{\Sigma_x^{1/2}\hat{\Phi}_{1:k}} - \mathbf{P}_{\Sigma_x^{1/2}\Phi_{1:k}} \right\|^2, \quad \left\| \mathbf{P}_{\Sigma_x^{1/2}\hat{\Phi}_{1:k}} - \mathbf{P}_{\Sigma_x^{1/2}\Phi_{1:k}} \right\|_F^2.$$

2.3 Measuring Subspace Distance by Principal Angles

In this section, we show that the excess prediction loss derived in (2.5) and (2.6) are closely related to the principal angles between the two subspaces spanned by the predictors. For any p dimensional random vector \mathbf{x} with mean zero and bounded second moments, define the Hilbert space

$$\mathcal{H} = \text{span}(\mathbf{x}) = \{X | X = \mathbf{x}^\top \mathbf{w}, \mathbf{w} \in \mathbb{R}^p\}$$

with covariance operator as the inner product, that is, for any $X_1, X_2 \in \mathcal{H}$, $\langle X_1, X_2 \rangle = \text{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2)$. For any pair of full column rank matrices $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{R}^{p \times k}$, consider the canonical correlation analysis between the two subspaces: $\text{span}(\mathbf{x}^\top \mathbf{U}_1)$ and $\text{span}(\mathbf{x}^\top \mathbf{U}_2)$. Suppose $(W_1, \widehat{W}_1), (W_2, \widehat{W}_2), \dots, (W_k, \widehat{W}_k)$ are the first, second, ..., k_{th} pair of canonical variates between $\text{span}(\mathbf{x}^\top \mathbf{U}_1)$ and $\text{span}(\mathbf{x}^\top \mathbf{U}_2)$, that is $\text{span}(W_1, \dots, W_k) = \text{span}(\mathbf{x}^\top \mathbf{U}_1)$, $\text{span}(\widehat{W}_1, \dots, \widehat{W}_k) = \text{span}(\mathbf{x}^\top \mathbf{U}_2)$ and $\langle W_i, W_j \rangle = \langle W_i, \widehat{W}_j \rangle = \langle \widehat{W}_i, \widehat{W}_j \rangle = 0$, for any $i \neq j$ and $\text{Var}(W_i) = \text{Var}(\widehat{W}_i) = 1$, for $i = 1, \dots, k$. Then $\{W_1, \dots, W_k\}$ and $\{\widehat{W}_1, \dots, \widehat{W}_k\}$ are orthonormal bases of $\text{span}(\mathbf{x}^\top \mathbf{U}_1)$ and $\text{span}(\mathbf{x}^\top \mathbf{U}_2)$, respectively. The i_{th} principal angle is defined as $\theta_i = \angle(W_i, \widehat{W}_i)$. Without loss of generality we assume $\theta_1 \geq \dots \geq \theta_k$ and define the ‘distance’ between the two subspaces as:

$$\mathcal{L}_2(\text{span}(\mathbf{x}^\top \mathbf{U}_1), \text{span}(\mathbf{x}^\top \mathbf{U}_2)) := \sum_{i=1}^k \sin^2 \theta_i = \sum_{i=1}^k \left(1 - \left| \langle W_i, \widehat{W}_i \rangle \right|^2 \right).$$

This is a valid ‘metric’ because the principal angles are uniquely defined while the canonical variates need not be. Since $\mathbf{x}^\top \boldsymbol{\Sigma}_x^{-1/2}$ is an orthonormal basis of \mathcal{H} under the covariance inner product, it is convenient to represent the elements in \mathcal{H} by this basis. Let

$$(W_1, \dots, W_k) = \mathbf{x}^\top \boldsymbol{\Sigma}_x^{-1/2} \mathbf{B}, \text{ and } (\widehat{W}_1, \dots, \widehat{W}_k) = \mathbf{x}^\top \boldsymbol{\Sigma}_x^{-1/2} \widehat{\mathbf{B}},$$

where $\mathbf{B} := [\mathbf{b}_1, \dots, \mathbf{b}_k]$, $\widehat{\mathbf{B}} := [\widehat{\mathbf{b}}_1, \dots, \widehat{\mathbf{b}}_k] \in \mathbb{R}^{p \times k}$ are the coordinate representations under $\mathbf{x}^\top \boldsymbol{\Sigma}_x^{-1/2}$. Then $\mathbf{B}, \widehat{\mathbf{B}}$ are $p \times k$ basis matrices. Moreover, we have $\mathbf{b}_i^\top \widehat{\mathbf{b}}_j = \langle W_i, \widehat{W}_j \rangle = 0$, for all $i \neq j$.

Let’s now see how to represent $\mathcal{L}_2(\text{span}(\mathbf{x}^\top \mathbf{U}_1), \text{span}(\mathbf{x}^\top \mathbf{U}_2))$ by \mathbf{B} and $\widehat{\mathbf{B}}$. In fact, since

$$1 - \left| \langle W_i, \widehat{W}_i \rangle \right|^2 = 1 - \left| \mathbf{b}_i^\top \widehat{\mathbf{b}}_i \right|^2 = \frac{1}{2} \left\| \mathbf{b}_i \mathbf{b}_i^\top - \widehat{\mathbf{b}}_i \widehat{\mathbf{b}}_i^\top \right\|_F^2,$$

we have

$$\begin{aligned} \mathcal{L}_2(\text{span}(\mathbf{x}^\top \mathbf{U}_1), \text{span}(\mathbf{x}^\top \mathbf{U}_2)) &= \frac{1}{2} \sum_{i=1}^k \left\| \mathbf{b}_i \mathbf{b}_i^\top - \widehat{\mathbf{b}}_i \widehat{\mathbf{b}}_i^\top \right\|_F^2 \\ &= \frac{1}{2} \left\| \sum_{i=1}^k (\mathbf{b}_i \mathbf{b}_i^\top - \widehat{\mathbf{b}}_i \widehat{\mathbf{b}}_i^\top) \right\|_F^2 \\ &= \frac{1}{2} \left\| \mathbf{B} \mathbf{B}^\top - \widehat{\mathbf{B}} \widehat{\mathbf{B}}^\top \right\|_F^2. \end{aligned}$$

Here the second equality is due to $\mathbf{b}_i^\top \mathbf{b}_j = \widehat{\mathbf{b}}_i^\top \widehat{\mathbf{b}}_j = \widehat{\mathbf{b}}_i^\top \mathbf{b}_j = 0$, for all $i \neq j$.

Finally, notice that $\text{span}(\mathbf{x}^\top \mathbf{U}_1) = \text{span}(W_1, \dots, W_k)$, $\mathbf{x}^\top \mathbf{U}_1 = (\mathbf{x}^\top \boldsymbol{\Sigma}_x^{-1/2})(\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_1)$, and $(W_1, \dots, W_k) = \mathbf{x}^\top \boldsymbol{\Sigma}_x^{-1/2} \mathbf{B}$. Then \mathbf{B} and $\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_1$ have the same column space. Since $\mathbf{B} \in \mathbb{R}^{p \times k}$ is a basis matrix, we have $\mathbf{B} \mathbf{B}^\top = \mathbf{P}_{\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_1}$, which is the orthogonal projector to the column space of $\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_1$. Similarly, we have $\widehat{\mathbf{B}} \widehat{\mathbf{B}}^\top = \mathbf{P}_{\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_2}$, which implies

$$\mathcal{L}_2(\text{span}(\mathbf{x}^\top \mathbf{U}_1), \text{span}(\mathbf{x}^\top \mathbf{U}_2)) = \frac{1}{2} \left\| \mathbf{P}_{\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_1} - \mathbf{P}_{\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_2} \right\|_F^2.$$

Similarly, we can also define the closeness through the leading principal angle:

$$\mathcal{L}_1(\text{span}(\mathbf{x}^\top \mathbf{U}_1), \text{span}(\mathbf{x}^\top \mathbf{U}_2)) := \sin^2 \theta_1 = 1 - \left| \langle W_1, \widehat{W}_1 \rangle \right|^2.$$

Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$, then $\mathbf{B} \mathbf{B}^\top \widehat{\mathbf{B}} \widehat{\mathbf{B}}^\top = \mathbf{B} \text{diag}(\cos(\boldsymbol{\theta})) \widehat{\mathbf{B}}^\top$, and by Lemma 5.8,

$$\begin{aligned} \left\| \mathbf{P}_{\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_1} - \mathbf{P}_{\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_2} \right\|^2 &= \left\| \mathbf{B} \mathbf{B}^\top - \widehat{\mathbf{B}} \widehat{\mathbf{B}}^\top \right\|^2 = 1 - \sigma_{\min}^2 \left(\mathbf{B} \mathbf{B}^\top \widehat{\mathbf{B}} \widehat{\mathbf{B}}^\top \right) \\ &= \sin^2(\theta_1) = \mathcal{L}_1(\text{span}(\mathbf{x}^\top \mathbf{U}_1), \text{span}(\mathbf{x}^\top \mathbf{U}_2)) \end{aligned}$$

We summarize the results into the following theorem.

Theorem 2.1 Suppose $(\mathbf{x}, z) \sim \mathbb{P}$ for some unknown distribution \mathbb{P} with covariance structure specified in (2.3) and subspace angles defined above. For any reduction matrices $\mathbf{U}, \mathbf{U}_\star \in \mathbb{R}^{p \times k}$,

$$\begin{aligned} \text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U})) - \text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U}_\star)) &= \sigma_z^2 \left(\mathbf{r}_{xz}^\top \left(\mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_\star} \right) \mathbf{r}_{xz} \right) \\ \sup_{\|\mathbf{r}_{xz}\|^2 = R^2} \text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U})) - \text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U}_\star)) & \\ &= \sigma_z^2 R^2 \left\| \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_\star} \right\| = \sigma_z^2 R^2 \|\sin(\boldsymbol{\theta})\|_\infty. \end{aligned}$$

By treating \mathbf{U}_\star as an oracle reduction matrix, let $\mathcal{A} = \{\mathbf{r} : \|\mathbf{r}\|^2 = R^2, \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_\star} \mathbf{r} = \mathbf{r}\}$

$$\begin{aligned} \sup_{\mathbf{r}_{xz} \in \mathcal{A}} \{ \text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U})) - \text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U}_\star)) \} & \\ &= \sigma_z^2 R^2 \left\| \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_\star} \right\|^2 = \sigma_z^2 R^2 \|\sin(\boldsymbol{\theta})\|_\infty^2. \end{aligned}$$

By treating \mathbf{U}_\star as a Bayes oracle, that is $\mathbf{r}_{xz} \sim \pi$ where π is the uniform measure (Haar measure) on \mathcal{A} , then

$$\begin{aligned} \mathbb{E}_{\mathbf{r}_{xz} \sim \pi} \{ \text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U})) - \text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U}_\star)) \} & \\ &= \frac{\sigma_z^2 R^2}{2k} \left\| \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_\star} \right\|_F^2 = \frac{\sigma_z^2 R^2}{k} \|\sin(\boldsymbol{\theta})\|_2^2 \end{aligned}$$

2.4 CCA for Multi-view Dimension Reduction

In the research and applications of multi-media analytics, data of the same object, is collected from multiple sources and exhibit heterogeneous properties. Features obtained from different domains are referred to as different ‘views’. Usually each view summarizes a specific aspect of the studied object and different views are complementary to one another. For example, in web-page classification, the hyperlink structure and the words on the page are two different views [Chaudhuri et al., 2009]. In video surveillance, images of cameras from different angles constitute different views [Loy et al., 2009]. For more recent results, see the survey paper of Xu et al. [2013] and references therein.

Although multiple views provide more potential discriminative information to distinguish the patterns of different classes, the feature vector of each view usually lies in a high dimensional space. It is critical, both statistically and computationally, to perform dimension reduction before applying any supervised learning algorithm. It has been shown by many researchers that canonical correlation analysis can achieve sufficient dimension reduction under certain multi-view assumptions [Kakade and Foster, 2007; Foster et al., 2008; Sridharan and Kakade, 2008; Fukumizu et al., 2009; Chaudhuri et al., 2009].

Suppose the input variable \mathbf{x} can be split into two views $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ and the goal is to predict the response z based on the two views. Let $(\boldsymbol{\Phi}_{1:k}, \boldsymbol{\Psi}_{1:k})$ be the top- k population canonical coefficients between $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$. With notations in this paper, Foster et al. [2008] essentially proved the following theorem.

Theorem 2.2 (*Sufficient Multi-view Dimension Reduction by CCA*) Under certain multi-view assumptions¹,

$$\text{loss}(z|\mathbf{x}^{(1)}) = \text{loss}(z|\text{span}((\mathbf{x}^{(1)})^\top \boldsymbol{\Phi}_{1:k})), \quad \text{loss}(z|\mathbf{x}^{(2)}) = \text{loss}(z|\text{span}((\mathbf{x}^{(2)})^\top \boldsymbol{\Psi}_{1:k}))$$

The theorem shows that all the predictive power of the original high-dimensional predictors $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ is fully captured by the top k canonical variates. However, the theorem focuses on the population level and does not take into account the estimation error induced by replacing the population canonical coefficients with the sample estimates. Such sample-population discrepancy can be quantified by

$$\begin{aligned} & \text{loss}(z|\text{span}((\mathbf{x}^{(1)})^\top \hat{\boldsymbol{\Phi}}_{1:k})) - \text{loss}(z|\text{span}((\mathbf{x}^{(1)})^\top \boldsymbol{\Phi}_{1:k})), \\ & \text{loss}(z|\text{span}((\mathbf{x}^{(2)})^\top \hat{\boldsymbol{\Psi}}_{1:k})) - \text{loss}(z|\text{span}((\mathbf{x}^{(2)})^\top \boldsymbol{\Psi}_{1:k})). \end{aligned}$$

which are exactly characterized by the proposed loss functions according to Theorem 2.1.

3 Theory

In this section, we introduce our main results on non-asymptotic upper and lower bounds for estimating CCA under the proposed loss functions. Specifically, the upper bound is achieved by sample CCA.

3.1 Reduction for Sample CCA

The linear invariance of both population canonical variates and sample canonical variates enables us to reduce the estimation error of sample CCA to the special case where $\boldsymbol{\Sigma}_x = \mathbf{I}_{p_1}$, $\boldsymbol{\Sigma}_y = \mathbf{I}_{p_2}$ and $\boldsymbol{\Sigma}_{xy} = [\boldsymbol{\Lambda} \mathbf{0}]$ (we assume $p_1 \leq p_2$ without loss of generality).

Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ be the data matrix where $\mathbf{x}_1, \dots, \mathbf{x}_n$ are *i.i.d.* observations from some distribution with covariance matrix $\boldsymbol{\Sigma}_x$ and similarly we define \mathbf{Y} . It is well known that the sample CCA can be formulated as the solution to the following optimization problem

$$\begin{aligned} (\hat{\boldsymbol{\Phi}}_{1:k}, \hat{\boldsymbol{\Psi}}_{1:k}) &= \arg \max_{\mathbf{W}_x, \mathbf{W}_y} \text{tr}(\mathbf{W}_x^\top \hat{\boldsymbol{\Sigma}}_{xy} \mathbf{W}_y) \\ \text{subject to } & \mathbf{W}_x^\top \hat{\boldsymbol{\Sigma}}_x \mathbf{W}_x = \mathbf{I}_k, \quad \mathbf{W}_y^\top \hat{\boldsymbol{\Sigma}}_y \mathbf{W}_y = \mathbf{I}_k. \end{aligned} \tag{3.1}$$

where $\hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y, \hat{\boldsymbol{\Sigma}}_{xy}$ are sample covariance matrices defined as

$$\hat{\boldsymbol{\Sigma}}_{xy} = \frac{1}{n} \mathbf{X}^\top \mathbf{Y}, \quad \hat{\boldsymbol{\Sigma}}_x = \frac{1}{n} \mathbf{X}^\top \mathbf{X}, \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_y = \frac{1}{n} \mathbf{Y}^\top \mathbf{Y}.$$

In the original definition, the loadings $\boldsymbol{\Psi}$ is a $p_2 \times p_1$ matrix. In this subsection, we abuse notation and redefine $\boldsymbol{\Psi}$ as a $p_2 \times p_2$ matrix by adding $p_2 - p_1$ columns such that $\boldsymbol{\Sigma}_y^{1/2} \boldsymbol{\Psi} \in \mathcal{O}(p_2)$. Let $\mathbf{a}_i = \boldsymbol{\Phi}^\top \mathbf{x}_i$ and $\mathbf{b}_i = \boldsymbol{\Psi}^\top \mathbf{y}_i$. Notice that $\boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_x \boldsymbol{\Phi} = \mathbf{I}_{p_1}$ and $\boldsymbol{\Psi}^\top \boldsymbol{\Sigma}_y \boldsymbol{\Psi} = \mathbf{I}_{p_2}$. Then we have $\mathbf{a}_i \stackrel{i.i.d.}{\sim} \mathbf{a}$ with distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_{p_1})$ and $\mathbf{b}_i \stackrel{i.i.d.}{\sim} \mathbf{b}$ with distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_{p_2})$. Moreover, we have

$$\boldsymbol{\Sigma}_{ab} := \mathbb{E} \mathbf{a}_i \mathbf{b}_i^\top = \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_{xy} \boldsymbol{\Psi} = \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_x \boldsymbol{\Phi} \boldsymbol{\Lambda} \boldsymbol{\Psi}^\top \boldsymbol{\Sigma}_y \boldsymbol{\Psi} = [\boldsymbol{\Lambda} \mathbf{0}].$$

¹See Foster et al. [2008] for details

This implies that

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \stackrel{i.i.d}{\sim} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{I}_{p_1} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \mathbf{I}_{p_2} \end{bmatrix}\right),$$

and $(\boldsymbol{\Phi}_{1:k}^a, \boldsymbol{\Psi}_{1:k}^a) = (\mathbf{I}_{p_1,k}, \mathbf{I}_{p_2,k})$ is the population top- k CCA pairs for (\mathbf{a}, \mathbf{b}) . Let $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_n)^\top = \mathbf{X}\boldsymbol{\Phi}$ and $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_n)^\top = \mathbf{Y}\boldsymbol{\Psi}$. Then

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_a &= \frac{1}{n} \mathbf{A}^\top \mathbf{A} = \boldsymbol{\Phi}^\top \hat{\boldsymbol{\Sigma}}_x \boldsymbol{\Phi}, \\ \hat{\boldsymbol{\Sigma}}_b &= \frac{1}{n} \mathbf{B}^\top \mathbf{B} = \boldsymbol{\Psi}^\top \hat{\boldsymbol{\Sigma}}_y \boldsymbol{\Psi}, \\ \hat{\boldsymbol{\Sigma}}_{ab} &= \frac{1}{n} \mathbf{A}^\top \mathbf{B} = \boldsymbol{\Phi}^\top \hat{\boldsymbol{\Sigma}}_{xy} \boldsymbol{\Psi}. \end{aligned}$$

Since $(\hat{\boldsymbol{\Phi}}_{1:k}, \hat{\boldsymbol{\Psi}}_{1:k})$ is a solution to the sample CCA (3.1), we know that $(\boldsymbol{\Phi}^{-1} \hat{\boldsymbol{\Phi}}_{1:k}, \boldsymbol{\Psi}^{-1} \hat{\boldsymbol{\Psi}}_{1:k})$ must be a solution to

$$\begin{aligned} & \max_{\mathbf{W}_x, \mathbf{W}_y} \text{tr}(\mathbf{W}_x^\top \boldsymbol{\Phi}^\top \hat{\boldsymbol{\Sigma}}_{xy} \boldsymbol{\Psi} \mathbf{W}_y) \\ \text{subject to } & \mathbf{W}_x^\top \boldsymbol{\Phi}^\top \hat{\boldsymbol{\Sigma}}_x \boldsymbol{\Phi} \mathbf{W}_x = \mathbf{I}_k, \quad \mathbf{W}_y^\top \boldsymbol{\Psi}^\top \hat{\boldsymbol{\Sigma}}_y \boldsymbol{\Psi} \mathbf{W}_y = \mathbf{I}_k, \end{aligned}$$

which is equivalent to

$$\begin{aligned} & \max_{\mathbf{W}_x, \mathbf{W}_y} \text{tr}(\mathbf{W}_x^\top \hat{\boldsymbol{\Sigma}}_{ab} \mathbf{W}_y) \\ \text{subject to } & \mathbf{W}_x^\top \hat{\boldsymbol{\Sigma}}_a \mathbf{W}_x = \mathbf{I}_k, \quad \mathbf{W}_y^\top \hat{\boldsymbol{\Sigma}}_b \mathbf{W}_y = \mathbf{I}_k. \end{aligned}$$

Therefore, $(\boldsymbol{\Phi}^{-1} \hat{\boldsymbol{\Phi}}_{1:k}, \boldsymbol{\Psi}^{-1} \hat{\boldsymbol{\Psi}}_{1:k})$ is indeed the sample canonical vectors for (\mathbf{a}, \mathbf{b}) , which we denote by $(\hat{\boldsymbol{\Phi}}_{1:k}^a, \hat{\boldsymbol{\Psi}}_{1:k}^b)$. By the discussion in Section 2.3,

$$\begin{aligned} \left\| P_{\boldsymbol{\Sigma}_x^{1/2} \boldsymbol{\Phi}_{1:k}} - P_{\boldsymbol{\Sigma}_x^{1/2} \hat{\boldsymbol{\Phi}}_{1:k}} \right\|^2 &= \mathcal{L}_1(\text{span}(\mathbf{x}^\top \hat{\boldsymbol{\Phi}}_{1:k}), \text{span}(\mathbf{x}^\top \boldsymbol{\Phi}_{1:k})) \\ &= \mathcal{L}_1(\text{span}(\mathbf{a}^\top \hat{\boldsymbol{\Phi}}_{1:k}^a), \text{span}(\mathbf{a}^\top \boldsymbol{\Phi}_{1:k}^a)) \\ &= \left\| P_{\boldsymbol{\Phi}_{1:k}^a} - P_{\hat{\boldsymbol{\Phi}}_{1:k}^a} \right\|^2 \end{aligned}$$

By the same argument,

$$\left\| P_{\boldsymbol{\Sigma}_x^{1/2} \boldsymbol{\Phi}_{1:k}} - P_{\boldsymbol{\Sigma}_x^{1/2} \hat{\boldsymbol{\Phi}}_{1:k}} \right\|_F^2 = \left\| P_{\boldsymbol{\Phi}_{1:k}^a} - P_{\hat{\boldsymbol{\Phi}}_{1:k}^a} \right\|_F^2$$

To sum up, it suffices to consider the special covariance structure $\boldsymbol{\Sigma}_x = \mathbf{I}_{p_1}, \boldsymbol{\Sigma}_y = \mathbf{I}_{p_2}, \boldsymbol{\Sigma}_{xy} = [\boldsymbol{\Lambda} \mathbf{0}]$ for the estimation error of sample CCA.

Remark 3.1 *The reduction argument essentially reveals the fact that the estimation error of sample CCA is independent of the condition numbers of the covariance matrices: $\kappa(\boldsymbol{\Sigma}_x)$ and $\kappa(\boldsymbol{\Sigma}_y)$. It is worthwhile to mention that such phenomenon is due to the linear invariance of sample canonical variates and might not exist for other estimators.*

3.2 Upper and Lower Bounds

In this section, we assume $\mathbf{x} \in \mathbb{R}^{p_1}, \mathbf{y} \in \mathbb{R}^{p_2}$ are jointly normal with mean zero and joint covariance matrix Σ specified by

$$\Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^\top & \Sigma_y \end{bmatrix}$$

where Σ_x and Σ_y are nonsingular. Let $\lambda_1, \dots, \lambda_{p_1 \wedge p_2}$ be the canonical correlations and (Φ, Ψ) be the canonical coefficient matrices (loadings) as defined in (1.1). For any $1 \leq k < p_1 \wedge p_2$, define the k_{th} eigen-gap as $\Delta = \lambda_k - \lambda_{k+1}$.

Theorem 3.2 (*Upper bound*) *There exists universal positive constants C, C_1, C_2 independent of n, p_1, p_2 and Σ such that if $n \geq C(p_1 + p_2)$, the sample CCA coefficients $\hat{\Phi}_{1:k}$ satisfies*

$$\begin{aligned} \mathbb{E} \left[\left\| P_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|^2 \right] &\leq C_1 \frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{\Delta^2} \frac{p_1}{n} + C_2 \left(\frac{p_1 + p_2}{n\Delta^2} \right)^2 \\ \mathbb{E} \left[\left\| P_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|_F^2 / k \right] &\leq C_1 \frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{\Delta^2} \frac{p_1 - k}{n} + C_2 \left(\frac{p_1 + p_2}{n\Delta^2} \right)^2 \end{aligned}$$

Symmetric upper bounds hold for $\hat{\Psi}_{1:k}$ by switching p_1 and p_2 .

This theorem exhibits several notable features:

1. The multiplicative factor $(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)/\Delta^2$ appears in the principal term. The inverse dependence on the eigen-gap Δ^2 is inherent for spectral estimations. The factor $(1 - \lambda_k^2)$ reveals that the estimation error decreases with increasing correlations between the two sets of random variables. When there is perfect correlation, that is $\lambda_k = 1$, we can recover the CCA directions errorlessly because the observed data along those directions are perfectly co-linear. The factor $(1 - \lambda_{k+1}^2)$ comes as surprise and appears in our lower bound as well. When λ_k is close to 1 and the eigen-gap Δ is close to 0, for example, in the regime $\lambda_k = 1 - C\Delta$ and $\Delta \rightarrow 0$, then

$$(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)/\Delta^2 \asymp \text{constant}$$

which shows consistency can still be achieved. On the contrast, without the factor $(1 - \lambda_{k+1}^2)$, the error will explode since $(1 - \lambda_k^2)/\Delta^2 \asymp 1/\Delta$. We remark that λ_k, λ_{k+1} are parameters depending on the dimensions and should not be treated as constant. As the number of variables increases, one should expect the canonical correlations to increase as well.

To the best of our knowledge, this is the first time for $(1 - \lambda_k^2)$ and $(1 - \lambda_{k+1}^2)$ to be obtained as a finite sample result. This is achieved by Taylor expanding the estimating equations for $\hat{\Phi}_{1:k}$ and $\hat{\Psi}_{1:k}$, inspired by the classical multivariate theory of [Anderson \[1963, 1984, 1999\]](#), while the analysis of [Gao et al. \[2014, 2015\]](#) and the recent work by [Cai and Zhang \[2016\]](#) does not yield this factor.

2. The dimension parameter p_2 only appears in the high order term, which implies that even in the regime $p_2 \gg p_1$, as long as the sample size is large enough (see Corollary 3.5), the estimation error of $\hat{\Phi}_{1:k}$ does not depend on p_2 . This phenomenon was first revealed in Gao et al. [2015] through multi-stage estimation and sample splitting. The recent work of Cai and Zhang [2016] directly proved such kind of result for sample CCA without splitting the samples. The results of both Gao et al. [2015] and Cai and Zhang [2016] are based on the extra assumption that all the residual canonical correlations are zero: $\lambda_{k+1} = \dots = \lambda_{p_1} = 0$ (or equivalently, the rank of Σ_{xy} is k).
3. The upper bound does not depend on the condition number of the covariance matrices: $\kappa(\Sigma_x)$ and $\kappa(\Sigma_y)$. It is directly implied by the reduction argument but not obvious because the separate estimation errors, $\|\Sigma_x - \hat{\Sigma}_x\|, \|\Sigma_y - \hat{\Sigma}_y\|, \|\Sigma_{xy} - \hat{\Sigma}_{xy}\|$, are in fact proportional to these condition numbers. The success of the reduction argument relies on the linear invariance of both population and sample canonical variates. For loss functions directly based on the loadings [Chen et al., 2013; Gao et al., 2014], the upper bounds will be proportional to these condition numbers.
4. The only assumption made in Theorem 3.2 is that Σ_x, Σ_y are invertible. Anderson [1999] assumes the canonical correlations are all distinct because the argument requires the asymptotic convergence of each individual canonical coefficient. Moreover, the result is asymptotic without finite sample guarantee. Gao et al. [2015]; Cai and Zhang [2016] assume $\lambda_{k+1} = \dots = \lambda_{p_1} = 0$ and Gao et al. [2014, 2015] assume the condition number $\kappa(\Sigma_x)$ and $\kappa(\Sigma_y)$ are bounded.

To establish the minimax lower bound, we define the parameter space $\mathcal{F}(p_1, p_2, k, \lambda_k, \lambda_{k+1}, \kappa_1, \kappa_2)$ as the collection of joint covariance matrices Σ satisfying

$$\Sigma_x, \Sigma_y \text{ are nonsingular, } \kappa(\Sigma_x) = \kappa_1, \kappa(\Sigma_y) = \kappa_2$$

$$0 \leq \lambda_{p_1 \wedge p_2} \leq \dots \leq \lambda_{k+1} < \lambda_k \leq \dots \leq \lambda_1 \leq 1$$

We deliberately specify $\kappa(\Sigma_x) = \kappa_1, \kappa(\Sigma_y) = \kappa_2$ to show the lower bound is independent of the condition number as well. For the rest of the paper, we will use the shorthand \mathcal{F} to represent this parameter space for simplicity.

Theorem 3.3 (Lower bound) *There exists a universal constant c independent of n, p_1, p_2 and Σ such that*

$$\inf_{\hat{\Phi}_{1:k}} \sup_{\Sigma \in \mathcal{F}} \mathbb{E} \left[\left\| P_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|^2 \right] \geq c^2 \left\{ \left(\frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{\Delta^2} \frac{p_1 - k}{n} \right) \wedge 1 \wedge \frac{p_1 - k}{k} \right\}$$

$$\inf_{\hat{\Phi}_{1:k}} \sup_{\Sigma \in \mathcal{F}} \mathbb{E} \left[\left\| P_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|_F^2 / k \right] \geq c^2 \left\{ \left(\frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{\Delta^2} \frac{p_1 - k}{n} \right) \wedge 1 \wedge \frac{p_1 - k}{k} \right\}$$

where $\Delta = \lambda_k - \lambda_{k+1}$. The lower bounds for $\hat{\Psi}_{1:k}$ can be obtained by switching p_1 and p_2 .

The proof of the lower bound can be found in Section 5.8, which relies on Lemma 5.3 and constructing a hypothesis class based on this lemma.

Remark 3.4 *The upper and lower bound together shows that the condition number of Σ_x and Σ_y is neither cursing nor blessing.*

Gao et al. [2015] obtained minimax lower bounds for sparse CCA in high dimensional regime. Rephrasing their results without sparsity assumptions, they essentially proved

$$\mathbb{E} \left\{ \inf_{\mathbf{Q} \in \mathcal{O}(p_1)} \mathbb{E} \left[\left\| \mathbf{x}^\top \Phi_{1:k} - \mathbf{x}^\top \hat{\Phi}_{1:k} \mathbf{Q} \right\|_F^2 / k \right] \right\} \geq c^2 \left\{ \left(\frac{1 - \lambda_k^2}{\kappa_1 \lambda_k^2} \frac{p_1 - k}{n} \right) \wedge 1 \wedge \frac{p_1 - k}{k} \right\},$$

with the parameter space $\lambda_{p_1 \wedge p_2} = \dots = \lambda_{k+1} = 0 < \lambda_k \leq \dots \leq \lambda_1 \leq 1$. The inner expectation is with respect to an independent sample \mathbf{x} and the outer expectation is with respect to the data from which $\hat{\Phi}_{1:k}$ is constructed. The inverse dependency on the condition number in their results can be removed by noticing that

$$\inf_{\mathbf{Q} \in \mathcal{O}(p_1)} \mathbb{E} \left[\left\| \mathbf{x}^\top \Phi_{1:k} - \mathbf{x}^\top \hat{\Phi}_{1:k} \mathbf{Q} \right\|_F^2 \right] \geq \frac{1}{2} \left\| P_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|_F^2.$$

and apply Theorem 3.3 with $\lambda_{k+1} = 0$.

Corollary 3.5 *When $p_1 \geq 2k$, there exists a universal positive constant c such that when*

$$\frac{p_1 + p_2}{n\Delta^2} \leq c \frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{(1 + p_2/p_1)}, \quad (3.2)$$

the minimax rates are characterized by

$$\begin{aligned} \inf_{\hat{\Phi}_{1:k}} \sup_{\Sigma \in \mathcal{F}} \mathbb{E} \left[\left\| P_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|^2 \right] &\asymp \frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{n\Delta^2} p_1, \\ \inf_{\hat{\Phi}_{1:k}} \sup_{\Sigma \in \mathcal{F}} \mathbb{E} \left[\left\| P_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|_F^2 / k \right] &\asymp \frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{n\Delta^2} p_1. \end{aligned}$$

Remark 3.6 *For consistency, we only need the left hand side of (3.2) to converge to zero while in order for the high order term to be dominated by the principal term, the left hand side of (3.2) is required to converge to zero faster than the right hand size.*

4 Upper Bound: Proof of Theorem 3.2

Throughout the proof, we assume $p_1 = p_2$ for the ease of presentation and the same argument works for any p_1 and p_2 at the cost of heavier notations (when $p_1 \neq p_2$, in the definition (4.4), $\Lambda_2, \hat{\Lambda}_2$ will be rectangular instead of square matrices. As a result, the subsequent places where Λ_2 appears in the current proof should be understood as either Λ_2 or Λ_2^\top according to the dimensions in the specific context). We will still use p_1, p_2 ($p_1 \leq p_2$) to denote the dimension of \mathbf{x} and \mathbf{y} separately such that the results will be interpretable when $p_1 \neq p_2$.

By the reduction argument in Section 3.1, it suffices to consider

$$\Sigma_x = I_{p_1}, \quad \Sigma_y = I_{p_2}, \quad \Sigma_{xy} = \Lambda$$

where $\mathbf{\Lambda} = \text{Diag}(\lambda_1, \lambda_2, \dots, \lambda_{p_1}) \in \mathbb{R}^{p_1 \times p_1}$ is the diagonal matrix with $1 \geq \lambda_1 \geq \dots \geq \lambda_{p_1} \geq 0$. Under this setup

$$\mathbf{\Phi}_{1:p_1} = \mathbf{I}_{p_1}, \quad \mathbf{\Psi}_{1:p_1} = \mathbf{I}_{p_2, p_1}.$$

and $\lambda_1, \lambda_2, \dots, \lambda_{p_1}$ are the canonical correlations. Then the error metric is reduced to

$$\left\| P_{\Sigma_x^{1/2} \hat{\mathbf{\Phi}}_{1:k}} - P_{\Sigma_x^{1/2} \mathbf{\Phi}_{1:k}} \right\| = \left\| P_{\hat{\mathbf{\Phi}}_{1:k}} - P_{\mathbf{\Phi}_{1:k}} \right\|$$

where $\|\cdot\|$ denotes either operator or Frobenius norm. Divide $\hat{\mathbf{\Phi}}_{1:k}$ into two blocks such that $\hat{\mathbf{\Phi}}_{1:k} = \begin{bmatrix} \hat{\mathbf{\Phi}}_{1:k}^u \\ \hat{\mathbf{\Phi}}_{1:k}^l \end{bmatrix}$ where $\hat{\mathbf{\Phi}}_{1:k}^u$ and $\hat{\mathbf{\Phi}}_{1:k}^l$ are the upper $k \times k$ and lower $(p_1 - k) \times k$ sub-matrices of $\hat{\mathbf{\Phi}}_{1:k}$ respectively. Let $\mathbf{U} = \begin{bmatrix} \hat{\mathbf{\Phi}}_{1:k}^u \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{p_1 \times k}$. Then $P_{\hat{\mathbf{\Phi}}_{1:k}} = P_{\mathbf{U}}$ and by Wedin's $\sin \theta$ law [Wedin, 1972], there exists universal constant C such that

$$\left\| P_{\hat{\mathbf{\Phi}}_{1:k}} - P_{\mathbf{U}} \right\|^2 \leq \frac{C \left\| \hat{\mathbf{\Phi}}_{1:k} - \mathbf{U} \right\|^2}{\left(\sigma_k(\hat{\mathbf{\Phi}}_{1:k}) - \sigma_{k+1}(\mathbf{U}) \right)^2} = \frac{C \left\| \hat{\mathbf{\Phi}}_{1:k}^l \right\|^2}{\sigma_k^2(\hat{\mathbf{\Phi}}_{1:k})} \quad (4.1)$$

The denominator in (4.1) is close to 1 with high probability when $n \geq C(p_1 + p_2)$ for some constant C . The remaining proof will focus on obtaining upper bounds for the numerator. The upper bound in terms of operator norm is involved which we will present detailed proof. The Frobenius norm bound can be obtained in a very similar (but simpler) manner of which the proof is only sketched. The proof mainly contains the following steps:

1. Express explicitly (principal term + high order term) each cell of the matrix $\hat{\mathbf{\Phi}}_{1:k}^l$ by Taylor expanding the estimating equations of sample CCA.
2. Derive two separate deterministic upper bounds for the principal part of $\|\hat{\mathbf{\Phi}}_{1:k}^l\|$. One of the bounds is tight when λ_{k+1} is bounded away from 1 and the other one is tight when λ_{k+1} is bounded away from 0.
3. Derive deterministic bound for the high order terms.
4. Put pieces together and obtain the desired upper bounds in Theorem 3.2.

Throughout the proof, the constants c, C, C_1, \dots might change from line to line.

4.1 Taylor Expansion for $\hat{\mathbf{\Phi}}_{1:k}^l$

Recall that $\hat{\mathbf{\Phi}} \in \mathbb{R}^{p_1 \times p_1}$, $\hat{\mathbf{\Psi}} \in \mathbb{R}^{p_2 \times p_1}$ are the sample canonical coefficients. By definition, the sample canonical coefficients can be written as the solution to the following two estimating equations,

$$\hat{\Sigma}_{xy} \hat{\mathbf{\Psi}} = \hat{\Sigma}_x \hat{\mathbf{\Phi}} \hat{\mathbf{\Lambda}} \quad (4.2)$$

$$\hat{\Sigma}_{yx} \hat{\mathbf{\Phi}} = \hat{\Sigma}_y \hat{\mathbf{\Psi}} \hat{\mathbf{\Lambda}} \quad (4.3)$$

Divide the matrices into blocks,

$$\hat{\Sigma}_x = \begin{bmatrix} \hat{\Sigma}_x^{11} & \hat{\Sigma}_x^{12} \\ \hat{\Sigma}_x^{21} & \hat{\Sigma}_x^{22} \end{bmatrix}, \quad \hat{\Sigma}_y = \begin{bmatrix} \hat{\Sigma}_y^{11} & \hat{\Sigma}_y^{12} \\ \hat{\Sigma}_y^{21} & \hat{\Sigma}_y^{22} \end{bmatrix}, \quad \hat{\Sigma}_{xy} = \begin{bmatrix} \hat{\Sigma}_{xy}^{11} & \hat{\Sigma}_{xy}^{12} \\ \hat{\Sigma}_{xy}^{21} & \hat{\Sigma}_{xy}^{22} \end{bmatrix}, \quad \hat{\Sigma}_{yx} = \begin{bmatrix} \hat{\Sigma}_{yx}^{11} & \hat{\Sigma}_{yx}^{12} \\ \hat{\Sigma}_{yx}^{21} & \hat{\Sigma}_{yx}^{22} \end{bmatrix}$$

where $\hat{\Sigma}_x^{11}, \hat{\Sigma}_y^{11}, \hat{\Sigma}_{xy}^{11}, \hat{\Sigma}_{yx}^{11}$ are $k \times k$ matrices. Similarly, we define

$$\Lambda = \begin{bmatrix} \Lambda_1 & \\ & \Lambda_2 \end{bmatrix}, \quad \hat{\Lambda} = \begin{bmatrix} \hat{\Lambda}_1 & \\ & \hat{\Lambda}_2 \end{bmatrix}, \quad (4.4)$$

where $\Lambda_1, \hat{\Lambda}_1$ are also $k \times k$ matrices. Finally, we define $\hat{\Psi}_{1:k}^u \in \mathbb{R}^{k \times k}$, $\hat{\Psi}_{1:k}^l \in \mathbb{R}^{(p_2-k) \times k}$ in the same way as $\hat{\Phi}_{1:k}^u, \hat{\Phi}_{1:k}^l$. With these notations, we can write down the lower left $(p_1 - k) \times k$ sub-matrix of (4.2) and (4.3) explicitly as

$$\hat{\Sigma}_{xy}^{21} \hat{\Psi}_{1:k}^u + \hat{\Sigma}_{xy}^{22} \hat{\Psi}_{1:k}^l = \hat{\Sigma}_x^{21} \hat{\Phi}_{1:k}^u \hat{\Lambda}_1 + \hat{\Sigma}_x^{22} \hat{\Phi}_{1:k}^l \hat{\Lambda}_1, \quad (4.5)$$

$$\hat{\Sigma}_{yx}^{21} \hat{\Phi}_{1:k}^u + \hat{\Sigma}_{yx}^{22} \hat{\Phi}_{1:k}^l = \hat{\Sigma}_y^{21} \hat{\Psi}_{1:k}^u \hat{\Lambda}_1 + \hat{\Sigma}_y^{22} \hat{\Psi}_{1:k}^l \hat{\Lambda}_1. \quad (4.6)$$

Similarly, the upper left $k \times k$ sub-matrix of (4.2) and (4.3) can be written explicitly as

$$\hat{\Sigma}_{xy}^{11} \hat{\Psi}_{1:k}^u + \hat{\Sigma}_{xy}^{12} \hat{\Psi}_{1:k}^l = \hat{\Sigma}_x^{11} \hat{\Phi}_{1:k}^u \hat{\Lambda}_1 + \hat{\Sigma}_x^{12} \hat{\Phi}_{1:k}^l \hat{\Lambda}_1, \quad (4.7)$$

$$\hat{\Sigma}_{yx}^{11} \hat{\Phi}_{1:k}^u + \hat{\Sigma}_{yx}^{12} \hat{\Phi}_{1:k}^l = \hat{\Sigma}_y^{11} \hat{\Psi}_{1:k}^u \hat{\Lambda}_1 + \hat{\Sigma}_y^{12} \hat{\Psi}_{1:k}^l \hat{\Lambda}_1. \quad (4.8)$$

Manipulate the terms in (4.7),

$$\Lambda_1 \hat{\Psi}_{1:k}^u + (\hat{\Sigma}_{xy}^{11} - \Lambda_1) \hat{\Psi}_{1:k}^u + \hat{\Sigma}_{xy}^{12} \hat{\Psi}_{1:k}^l = \hat{\Phi}_{1:k}^u \hat{\Lambda}_1 + (\hat{\Sigma}_x^{11} - I_k) \hat{\Phi}_{1:k}^u \hat{\Lambda}_1 + \hat{\Sigma}_x^{12} \hat{\Phi}_{1:k}^l \hat{\Lambda}_1.$$

Therefore,

$$\begin{aligned} \Lambda_1 \hat{\Psi}_{1:k}^u - \hat{\Phi}_{1:k}^u \Lambda_1 &= \hat{\Phi}_{1:k}^u (\hat{\Lambda}_1 - \Lambda_1) + (\hat{\Sigma}_x^{11} - I_k) \hat{\Phi}_{1:k}^u \hat{\Lambda}_1 + \hat{\Sigma}_x^{12} \hat{\Phi}_{1:k}^l \hat{\Lambda}_1 \\ &\quad - (\hat{\Sigma}_{xy}^{11} - \Lambda_1) \hat{\Psi}_{1:k}^u - \hat{\Sigma}_{xy}^{12} \hat{\Psi}_{1:k}^l := \delta_1. \end{aligned} \quad (4.9)$$

The equation essentially implies $\Lambda_1 \hat{\Psi}_{1:k}^u \approx \hat{\Phi}_{1:k}^u \Lambda_1$ because δ_1 will be proved to be a higher order term. Apply the same argument to (4.8), and we will obtain

$$\begin{aligned} \Lambda_1 \hat{\Phi}_{1:k}^u - \hat{\Psi}_{1:k}^u \Lambda_1 &= \hat{\Psi}_{1:k}^u (\hat{\Lambda}_1 - \Lambda_1) + (\hat{\Sigma}_y^{11} - I_k) \hat{\Psi}_{1:k}^u \hat{\Lambda}_1 + \hat{\Sigma}_y^{12} \hat{\Psi}_{1:k}^l \hat{\Lambda}_1 \\ &\quad - (\hat{\Sigma}_{yx}^{11} - \Lambda_1) \hat{\Phi}_{1:k}^u - \hat{\Sigma}_{yx}^{12} \hat{\Phi}_{1:k}^l := \delta_2. \end{aligned} \quad (4.10)$$

Similarly, massage the terms in (4.5),

$$\begin{aligned} \hat{\Sigma}_{xy}^{21} \hat{\Psi}_{1:k}^u + \Lambda_2 \hat{\Psi}_{1:k}^l + (\hat{\Sigma}_{xy}^{22} - \Lambda_2) \hat{\Psi}_{1:k}^l &= \hat{\Sigma}_x^{21} \hat{\Phi}_{1:k}^u \Lambda_1 + \hat{\Sigma}_x^{21} \hat{\Phi}_{1:k}^u (\hat{\Lambda}_1 - \Lambda_1) \\ &\quad + \hat{\Phi}_{1:k}^l \Lambda_1 + (\hat{\Sigma}_x^{22} \hat{\Phi}_{1:k}^l \hat{\Lambda}_1 - \hat{\Phi}_{1:k}^l \Lambda_1), \end{aligned}$$

which can be equivalently written as

$$\begin{aligned} \hat{\Sigma}_{xy}^{21} \hat{\Psi}_{1:k}^u + \Lambda_2 \hat{\Psi}_{1:k}^l - \hat{\Sigma}_x^{21} \hat{\Phi}_{1:k}^u \Lambda_1 - \hat{\Phi}_{1:k}^l \Lambda_1 \\ = \hat{\Sigma}_x^{21} \hat{\Phi}_{1:k}^u (\hat{\Lambda}_1 - \Lambda_1) + (\hat{\Sigma}_x^{22} \hat{\Phi}_{1:k}^l \hat{\Lambda}_1 - \hat{\Phi}_{1:k}^l \Lambda_1) - (\hat{\Sigma}_{xy}^{22} - \Lambda_2) \hat{\Psi}_{1:k}^l := \delta_3. \end{aligned} \quad (4.11)$$

Apply the same argument to (4.6),

$$\begin{aligned} & \hat{\Sigma}_{yx}^{21} \hat{\Phi}_{1:k}^u + \Lambda_2 \hat{\Phi}_{1:k}^l - \hat{\Sigma}_y^{21} \hat{\Psi}_{1:k}^u \Lambda_1 - \hat{\Psi}_{1:k}^l \Lambda_1 \\ &= \hat{\Sigma}_y^{21} \hat{\Psi}_{1:k}^u (\hat{\Lambda}_1 - \Lambda_1) + (\hat{\Sigma}_y^{22} \hat{\Psi}_{1:k}^l \hat{\Lambda}_1 - \hat{\Psi}_{1:k}^l \Lambda_1) - (\hat{\Sigma}_{yx}^{22} - \Lambda_2) \hat{\Phi}_{1:k}^l := \delta_4. \end{aligned} \quad (4.12)$$

Consider (4.11) $\times (-\Lambda_1) - \Lambda_2 \times$ (4.12), then

$$\begin{aligned} & \hat{\Phi}_{1:k}^l \Lambda_1^2 - \Lambda_2^2 \hat{\Phi}_{1:k}^l + \hat{\Sigma}_x^{21} \hat{\Phi}_{1:k}^u \Lambda_1^2 - \hat{\Sigma}_{xy}^{21} \hat{\Psi}_{1:k}^u \Lambda_1 - \Lambda_2 \hat{\Sigma}_{yx}^{21} \hat{\Phi}_{1:k}^u + \Lambda_2 \hat{\Sigma}_y^{21} \hat{\Psi}_{1:k}^u \Lambda_1 \\ &= -(\delta_3 \Lambda_1 + \Lambda_2 \delta_4) := \delta_5, \end{aligned}$$

that is

$$\hat{\Phi}_{1:k}^l \Lambda_1^2 - \Lambda_2^2 \hat{\Phi}_{1:k}^l = \hat{\Sigma}_{xy}^{21} \hat{\Psi}_{1:k}^u \Lambda_1 + \Lambda_2 \hat{\Sigma}_{yx}^{21} \hat{\Phi}_{1:k}^u - \hat{\Sigma}_x^{21} \hat{\Phi}_{1:k}^u \Lambda_1^2 - \Lambda_2 \hat{\Sigma}_y^{21} \hat{\Psi}_{1:k}^u \Lambda_1 + \delta_5 \quad (4.13)$$

The equation above indicates that, ignoring $\hat{\Phi}_{1:k}^u, \hat{\Psi}_{1:k}^u$ and the high order term δ_5 , the target is expressed as a linear function of the sample covariance matrices. Later we will show that the sample covariance matrices and $\hat{\Phi}_{1:k}^u, \hat{\Psi}_{1:k}^u$ can be decoupled.

4.2 Upper Bounds for $\|\hat{\Phi}_{1:k}^l\|$

The proof in this part heavily relies on the following lemma about the Hadamard operator norm for some structured matrices.

Lemma 4.1 (*Hadamard Operator Norm*) For $A \in \mathbb{R}^{m \times n}$, define the Hadamard operator norm as

$$\|A\| = \sup \{ \|A \circ B\| : \|B\| \leq 1, B \in \mathbb{R}^{m \times n} \}$$

Let $\alpha_1, \dots, \alpha_m$ and β_1, \dots, β_n be arbitrary positive numbers lower bounded by a positive constant δ . Define $A_1, A_2, A_3 \in \mathbb{R}^{m \times n}$ by

$$[A_1]_{ij} = \frac{1}{\alpha_i + \beta_j}, [A_2]_{ij} = \frac{\min\{\alpha_i, \beta_j\}}{\alpha_i + \beta_j}, [A_3]_{ij} = \frac{\max\{\alpha_i, \beta_j\}}{\alpha_i + \beta_j}$$

Then

$$\|A_1\| \leq \frac{1}{2\delta}, \|A_2\| \leq \frac{1}{2}, \|A_3\| \leq \frac{3}{2}$$

See Section 5.2 for the proof of this lemma.

4.2.1 Upper Bound I: tight for $\lambda_{k+1} \leq 1/2$

Multiply both sides of (4.9) by Λ_1 on the right will yield

$$\Lambda_1 \hat{\Psi}_{1:k}^u \Lambda_1 - \hat{\Phi}_{1:k}^u \Lambda_1^2 = \delta_1 \Lambda_1. \quad (4.14)$$

Substitute (4.9), (4.10) and (4.14) into (4.13),

$$\begin{aligned}
\hat{\Phi}_{1:k}^l \Lambda_1^2 - \Lambda_2^2 \hat{\Phi}_{1:k}^l &= \hat{\Sigma}_{xy}^{21} \hat{\Psi}_{1:k}^u \Lambda_1 + \Lambda_2 \hat{\Sigma}_{yx}^{21} \hat{\Phi}_{1:k}^u - \hat{\Sigma}_x^{21} \Lambda_1 \hat{\Psi}_{1:k}^u \Lambda_1 + \hat{\Sigma}_x^{21} \delta_1 \Lambda_1 \\
&\quad - \Lambda_2 \hat{\Sigma}_y^{21} \Lambda_1 \hat{\Phi}_{1:k}^u - \Lambda_2 \hat{\Sigma}_y^{21} \delta_1 + \delta_5 \\
&= (\hat{\Sigma}_{xy}^{21} - \hat{\Sigma}_x^{21} \Lambda_1) \hat{\Psi}_{1:k}^u \Lambda_1 + \Lambda_2 (\hat{\Sigma}_{yx}^{21} - \hat{\Sigma}_y^{21} \Lambda_1) \hat{\Phi}_{1:k}^u \\
&\quad + \hat{\Sigma}_x^{21} \delta_1 \Lambda_1 - \Lambda_2 \hat{\Sigma}_y^{21} \delta_1 + \delta_5 \\
&:= B_1 \hat{\Psi}_{1:k}^u \Lambda_1 + \Lambda_2 B_2 \hat{\Phi}_{1:k}^u + \delta_6.
\end{aligned} \tag{4.15}$$

where $B_1 = \hat{\Sigma}_{xy}^{21} - \hat{\Sigma}_x^{21} \Lambda_1$, $B_2 = \hat{\Sigma}_{yx}^{21} - \hat{\Sigma}_y^{21} \Lambda_1$ and

$$\delta_6 = \hat{\Sigma}_x^{21} \delta_1 \Lambda_1 - \Lambda_2 \hat{\Sigma}_y^{21} \delta_1 + \delta_5.$$

Further, define the matrix $(p_1 - k) \times k$ matrices A_1, A_2 by

$$[A_1]_{ij} = \frac{1}{\lambda_j + \lambda_{k+i}}, [A_2]_{ij} = \frac{1}{\lambda_j - \lambda_{k+i}}, \quad 1 \leq i \leq p_1 - k, 1 \leq j \leq k$$

Then we can rewrite (4.15) as

$$\begin{aligned}
\hat{\Phi}_{1:k}^l &= A_1 \circ A_2 \circ (B_1 \hat{\Psi}_{1:k}^u \Lambda_1) + A_1 \circ A_2 \circ (\Lambda_2 B_2 \hat{\Phi}_{1:k}^u) + A_1 \circ A_2 \circ \delta_6 \\
&= (A_1 \Lambda_1) \circ A_2 \circ (B_1 \hat{\Psi}_{1:k}^u) + (\Lambda_2 A_1) \circ A_2 \circ (B_2 \hat{\Phi}_{1:k}^u) + A_1 \circ A_2 \circ \delta_6,
\end{aligned}$$

where \circ denotes the matrix Hadamard (element-wise) product. Define $\alpha_j = \lambda_j, 1 \leq j \leq k$ and $\beta_i = \lambda_{k+i}, 1 \leq i \leq p_1 - k$, then

$$[A_1 \Lambda_1]_{ij} = \frac{\lambda_j}{\lambda_j + \lambda_{k+i}} = \frac{\max\{\alpha_j, \beta_i\}}{\alpha_j + \beta_i}, [A_2 A_1]_{ij} = \frac{\lambda_{k+i}}{\lambda_j + \lambda_{k+i}} = \frac{\min\{\alpha_j, \beta_i\}}{\alpha_j + \beta_i} \tag{4.16}$$

Therefore, by Lemma 4.1,

$$\begin{aligned}
\|(A_1 \Lambda_1) \circ A_2 \circ (B_1 \hat{\Psi}_{1:k}^u)\| &\leq \frac{3}{2} \|A_2 \circ (B_1 \hat{\Psi}_{1:k}^u)\|, \\
\|(\Lambda_2 A_1) \circ A_2 \circ (B_2 \hat{\Phi}_{1:k}^u)\| &\leq \frac{1}{2} \|A_2 \circ (B_2 \hat{\Phi}_{1:k}^u)\|.
\end{aligned}$$

Observe that

$$[A_2]_{ij} = \frac{1}{\lambda_j - \lambda_{k+i}} = \frac{1}{(\lambda_j - \lambda_k + \Delta/2) + (\lambda_k - \Delta/2 - \lambda_{k+i})} = \frac{1}{a_j + b_i}, \tag{4.17}$$

where $a_j = \lambda_j - (\lambda_k - \Delta/2), 1 \leq j \leq k$ and $b_i = (\lambda_k - \Delta/2) - \lambda_{k+i}, 1 \leq i \leq p_1 - k$. Then $a_j, b_i \geq \Delta/2$ and again apply Lemma 4.1,

$$\begin{aligned}
\|A_2 \circ (B_1 \hat{\Psi}_{1:k}^u)\| &\leq \frac{1}{\Delta} \|B_1 \hat{\Psi}_{1:k}^u\| \leq \frac{1}{\Delta} \|B_1\| \|\hat{\Psi}_{1:k}^u\|, \\
\|A_2 \circ (B_2 \hat{\Phi}_{1:k}^u)\| &\leq \frac{1}{\Delta} \|B_2 \hat{\Phi}_{1:k}^u\| \leq \frac{1}{\Delta} \|B_2\| \|\hat{\Phi}_{1:k}^u\|.
\end{aligned}$$

which implies that

$$\|\hat{\Phi}_{1:k}^l\| \leq \frac{1}{2\Delta} \left(3 \|B_1\| \|\hat{\Psi}_{1:k}^u\| + \|B_2\| \|\hat{\Phi}_{1:k}^u\| \right) + \|A_1 \circ A_2 \circ \delta_6\| \tag{4.18}$$

In Section 4.4, we will show that this bound is tight and matches Theorem 3.2 when λ_{k+1} is away from 1. Now we switch to deriving the other upper bound which will be tight when λ_{k+1} is close to 1.

4.2.2 Upper Bound II: tight for $\lambda_{k+1} \geq 1/2$

Notice that $\mathbf{\Lambda}_1 \times (4.9) + \mathbf{\Lambda}_1 \times (4.10)$ yields

$$\mathbf{\Lambda}_1^2 \hat{\Psi}_{1:k}^u - \hat{\Psi}_{1:k}^u \mathbf{\Lambda}_1^2 = \mathbf{\Lambda}_1 \delta_1 + \delta_2 \mathbf{\Lambda}_1. \quad (4.19)$$

Substitute (4.9), (4.10) and (4.19) into (4.13),

$$\begin{aligned} \hat{\Phi}_{1:k}^l \mathbf{\Lambda}_1^2 - \mathbf{\Lambda}_1^2 \hat{\Phi}_{1:k}^l &= \hat{\Sigma}_{xy}^{21} \mathbf{\Lambda}_1 \hat{\Phi}_{1:k}^u + \hat{\Sigma}_{xy}^{21} \delta_1 + \mathbf{\Lambda}_2 \hat{\Sigma}_{yx}^{21} \hat{\Phi}_{1:k}^u - \hat{\Sigma}_x^{21} \mathbf{\Lambda}_1^2 \hat{\Phi}_{1:k}^u \\ &\quad + \hat{\Sigma}_x^{21} (\mathbf{\Lambda}_1 \delta_1 + \delta_2 \mathbf{\Lambda}_1) - \mathbf{\Lambda}_2 \hat{\Sigma}_y^{21} \mathbf{\Lambda}_1 \hat{\Phi}_{1:k}^u + \mathbf{\Lambda}_2 \hat{\Sigma}_y^{21} \delta_2 + \delta_5 \\ &= B \hat{\Phi}_{1:k}^u + \delta_7, \end{aligned} \quad (4.20)$$

where we define

$$\begin{aligned} B &= \hat{\Sigma}_{xy}^{21} \mathbf{\Lambda}_1 + \mathbf{\Lambda}_2 \hat{\Sigma}_{yx}^{21} - \hat{\Sigma}_x^{21} \mathbf{\Lambda}_1^2 - \mathbf{\Lambda}_2 \hat{\Sigma}_y^{21} \mathbf{\Lambda}_1, \\ \delta_7 &= \hat{\Sigma}_{xy}^{21} \delta_1 + \hat{\Sigma}_x^{21} (\mathbf{\Lambda}_1 \delta_1 + \delta_2 \mathbf{\Lambda}_1) + \mathbf{\Lambda}_2 \hat{\Sigma}_y^{21} \delta_2 + \delta_5. \end{aligned}$$

Again with the definition of \mathbf{A}_1 and \mathbf{A}_2 ,

$$\hat{\Phi}_{1:k}^l = \mathbf{A}_1 \circ \mathbf{A}_2 \circ (B \hat{\Phi}_{1:k}^u) + \mathbf{A}_1 \circ \mathbf{A}_2 \circ \delta_7.$$

Notice that we can rewrite \mathbf{A}_1 as

$$[\mathbf{A}_1]_{ij} = \frac{1}{\lambda_j + \lambda_{k+i}} = \frac{1}{(\lambda_j - \lambda_k/2) + (\lambda_{k+i} + \lambda_k/2)} = \frac{1}{\alpha_j + \beta_i},$$

where $\alpha_j = \lambda_j - \lambda_k/2, 1 \leq j \leq k$ and $\beta_i = \lambda_{k+i} + \lambda_k/2, 1 \leq i \leq p_1 - k$. Hence $\alpha_j, \beta_i \geq \lambda_k/2$, and apply Lemma 4.1,

$$\|\hat{\Phi}_{1:k}^l\| \leq \frac{1}{\lambda_k} \left(\|\mathbf{A}_2 \circ (B \hat{\Phi}_{1:k}^u)\| + \|\mathbf{A}_2 \circ \delta_7\| \right). \quad (4.21)$$

Define k^* as the largest index i such that

$$k+1 \leq i \leq p_1, \quad \lambda_i \geq 2\lambda_k - 1 - \Delta.$$

Divide the indexes $1, \dots, p_1 - k$ into two sets:

$$\mathcal{I}_1 = \{i - k : k+1 \leq i \leq k^*\}, \quad \mathcal{I}_2 = \{i - k : k^* + 1 \leq i \leq p_1\},$$

and accordingly divide \mathbf{A}_2 and $B \hat{\Phi}_{1:k}^u$ into two blocks:

$$\mathbf{A}_2 = \begin{bmatrix} \mathbf{A}_2^{(1)} \\ \mathbf{A}_2^{(2)} \end{bmatrix}, \quad B \hat{\Phi}_{1:k}^u = \begin{bmatrix} B^{(1)} \hat{\Phi}_{1:k}^u \\ B^{(2)} \hat{\Phi}_{1:k}^u \end{bmatrix}.$$

where $\mathbf{A}_2^{(1)}, B^{(1)}$ corresponds to the rows indexed by \mathcal{I}_1 and $\mathbf{A}_2^{(2)}, B^{(2)}$ corresponds to the rows indexed by \mathcal{I}_2 . Then

$$\mathbf{A}_2 \circ (B \hat{\Phi}_{1:k}^u) = \begin{bmatrix} \mathbf{A}_2^{(1)} \circ (B^{(1)} \hat{\Phi}_{1:k}^u) \\ \mathbf{A}_2^{(2)} \circ (B^{(2)} \hat{\Phi}_{1:k}^u) \end{bmatrix},$$

and by triangle inequality,

$$\left\| \mathbf{A}_2 \circ (\mathbf{B} \hat{\Phi}_{1:k}^u) \right\| \leq \left\| \mathbf{A}_2^{(1)} \circ (\mathbf{B}^{(1)} \hat{\Phi}_{1:k}^u) \right\| + \left\| \mathbf{A}_2^{(2)} \circ (\mathbf{B}^{(2)} \hat{\Phi}_{1:k}^u) \right\|.$$

For the first part, by the same argument as in (4.17)

$$\left\| \mathbf{A}_2^{(1)} \circ (\mathbf{B}^{(1)} \hat{\Phi}_{1:k}^u) \right\| \leq \frac{1}{\Delta} \left\| \mathbf{B}^{(1)} \right\| \left\| \hat{\Phi}_{1:k}^u \right\|. \quad (4.22)$$

For the second part, let $\mathbf{D} = \text{diag}(\lambda_k - \Delta/2 - \lambda_{k^*+1}, \dots, \lambda_k - \Delta/2 - \lambda_{p_1}) \in \mathbb{R}^{(p_1-k^*) \times (p_1-k^*)}$. Then

$$\mathbf{A}_2^{(2)} \circ (\mathbf{B}^{(2)} \hat{\Phi}_{1:k}^u) = (\mathbf{D} \mathbf{A}_2^{(2)}) \circ (\mathbf{D}^{-1} \mathbf{B}^{(2)} \hat{\Phi}_{1:k}^u).$$

Notice that for $1 \leq i \leq p_1 - k^*$, $1 \leq j \leq k$

$$\begin{aligned} [\mathbf{D} \mathbf{A}_2^{(2)}]_{ij} &= \frac{(\lambda_k - \Delta/2 - \lambda_{k^*+i})}{\lambda_j - \lambda_{i+k}} \\ &= \frac{(\lambda_k - \Delta/2 - \lambda_{k^*+i})}{\lambda_j - (\lambda_k - \Delta/2) + (\lambda_k - \Delta/2 - \lambda_{k^*+i})} \\ &= \frac{b_i}{a_j + b_i}, \end{aligned}$$

where $a_j = \lambda_j - (\lambda_k - \Delta/2)$, $1 \leq j \leq k$ and $b_i = (\lambda_k - \Delta/2) - \lambda_{k^*+i}$ for $1 \leq i \leq p_1 - k^*$. Further observe that by definition of k^* ,

$$b_i - a_j = (\lambda_k - \Delta/2) - \lambda_{k^*+i} - \lambda_j + (\lambda_k - \Delta/2) \geq 2\lambda_k - 1 - \Delta - \lambda_{k^*+i} \geq 0,$$

that is

$$[\mathbf{D} \mathbf{A}_2^{(2)}]_{ij} = \frac{\max\{a_j, b_i\}}{a_j + b_i}.$$

Again, by Lemma 4.1,

$$\left\| \mathbf{A}_2^{(2)} \circ (\mathbf{B}^{(2)} \hat{\Phi}_{1:k}^u) \right\| \leq \frac{3}{2} \left\| \mathbf{D}^{-1} \mathbf{B}^{(2)} \hat{\Phi}_{1:k}^u \right\| \leq \frac{3}{2} \left\| \mathbf{D}^{-1} \mathbf{B}^{(2)} \right\| \left\| \hat{\Phi}_{1:k}^u \right\|. \quad (4.23)$$

Substitute (4.22) and (4.23) into (4.21),

$$\left\| \hat{\Phi}_{1:k}^l \right\| \leq \frac{1}{\lambda_k} \left(\left\| \mathbf{A}_2 \circ (\mathbf{B} \hat{\Phi}_{1:k}^u) \right\| + \left\| \mathbf{A}_2 \circ \delta_7 \right\| \right) \quad (4.24)$$

$$\leq \frac{1}{2\lambda_k \Delta} \left\| \mathbf{B}^{(1)} \right\| \left\| \hat{\Phi}_{1:k}^u \right\| + \frac{3}{2\lambda_k} \left\| \mathbf{D}^{-1} \mathbf{B}^{(2)} \right\| \left\| \hat{\Phi}_{1:k}^u \right\| + \frac{1}{\lambda_k} \left\| \mathbf{A}_2 \circ \delta_7 \right\|. \quad (4.25)$$

4.3 Upper Bounds in Expectation for the Principal Terms

We state upper bounds for key quantities in (4.18) and (4.25) with proofs deferred to Section 5.

Lemma 4.2 *There exists universal constant C such that*

$$\mathbb{E}[\|\mathbf{B}_1\|^2], \mathbb{E}[\|\mathbf{B}_2\|^2] \leq C \frac{p_1}{n} (1 - \lambda_k^2). \quad (4.26)$$

See Section 5.3 for the proof of this lemma.

Lemma 4.3 *There exists universal constants c, C_1 such that the following inequality holds with probability at least $1 - 2e^{-ct^2}$,*

$$\|\mathbf{D}^{-1} \mathbf{B}^{(2)}\| \leq \sqrt{\frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{\Delta^2}} \max\{\delta, \delta^2\} \quad \delta = C_1 \left(\sqrt{\frac{p_1}{n}} + \frac{t}{\sqrt{n}} \right). \quad (4.27)$$

As a corollary, there exists constant C_2 ,

$$\mathbb{E}\|\mathbf{D}^{-1} \mathbf{B}^{(2)}\|^2 \leq C_2 \frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)p_1}{n\Delta^2}. \quad (4.28)$$

See Section 5.4 for the proof of this lemma.

Lemma 4.4 *There exists universal constants c, C_1 such that the following inequality holds with probability at least $1 - 2e^{-ct^2}$,*

$$\|\mathbf{B}^{(1)}\| \leq \sqrt{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)} \max\{\delta, \delta^2\} \quad \delta = C_1 \left(\sqrt{\frac{p_1}{n}} + \frac{t}{\sqrt{n}} \right). \quad (4.29)$$

As a corollary, there exists constant C_2 ,

$$\mathbb{E}\|\mathbf{B}^{(1)}\|^2 \leq C_2 \frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)p_1}{n}. \quad (4.30)$$

See Section 5.5 for the proof of this lemma.

4.4 Upper Bound for Operator Norm

For the ease of presentation, we introduce $\mathbf{z} = (\mathbf{x}^\top, \mathbf{y}^\top)^\top$ as the concatenation of \mathbf{x} and \mathbf{y} . Then the population and sample covariances of \mathbf{z} can be written as,

$$\boldsymbol{\Sigma}_z = \begin{bmatrix} \mathbf{I}_{p_1} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \mathbf{I}_{p_2} \end{bmatrix}, \quad \hat{\boldsymbol{\Sigma}}_z = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_x & \hat{\boldsymbol{\Sigma}}_{xy} \\ \hat{\boldsymbol{\Sigma}}_{yx} & \hat{\boldsymbol{\Sigma}}_y \end{bmatrix}.$$

The advantage of introducing \mathbf{z} is that the sample-population discrepancy for \mathbf{x} and \mathbf{y} can be simultaneously bounded by that of \mathbf{z} .

Lemma 4.5 *There exists universal constant C such that the following inequality holds deterministically,*

$$\|\mathbf{A}_1 \circ \mathbf{A}_2 \circ \boldsymbol{\delta}_6\|, \|\mathbf{A}_2 \circ \boldsymbol{\delta}_7\| \leq \frac{C\|\boldsymbol{\Sigma}_z - \hat{\boldsymbol{\Sigma}}_z\|^2}{\Delta^2} (2 + \|\hat{\boldsymbol{\Sigma}}_z\|)^2 (\|\hat{\boldsymbol{\Phi}}_{1:k}\| + \|\hat{\boldsymbol{\Psi}}_{1:k}\|) (1 + \|\boldsymbol{\Sigma}_z - \hat{\boldsymbol{\Sigma}}_z\|),$$

where $\Delta = \lambda_k - \lambda_{k+1}$ is the eigen-gap.

See Section 6.1 for the proof.

Lemma 4.6 *There exists universal constant c, C_1, C_2 such that when $n \geq C_1(p_1 + p_2)$, the following inequality holds*

$$\begin{aligned} \sigma_k(\hat{\boldsymbol{\Phi}}_{1:k})^2 &\geq 1/2, \\ (2 + \|\hat{\boldsymbol{\Sigma}}_z\|)^2 (\|\hat{\boldsymbol{\Phi}}_{1:k}\| + \|\hat{\boldsymbol{\Psi}}_{1:k}\|) (1 + \|\boldsymbol{\Sigma}_z - \hat{\boldsymbol{\Sigma}}_z\|) &\leq C_2, \end{aligned}$$

with probability at least $1 - e^{-cn}$.

See Section 5.6 for the proof.

Let G be the event that the inequalities in Lemma 4.6 hold. Notice that $\|\mathbf{P}_1 - \mathbf{P}_2\| \leq 1$ for any pair of projection matrices with equal rank. Substitute into equation (4.1),

$$\begin{aligned} \mathbb{E} \left\| \mathbf{P}_{\hat{\boldsymbol{\Phi}}_{1:k}} - \mathbf{P}_{\boldsymbol{\Phi}_{1:k}} \right\|^2 &\leq \mathbb{P}(G^c) + \mathbb{E} \left[\left\| \mathbf{P}_{\hat{\boldsymbol{\Phi}}_{1:k}} - \mathbf{P}_{\boldsymbol{\Phi}_{1:k}} \right\|^2 I_G \right] \\ &\leq \mathbb{P}(G^c) + \mathbb{E} \left[\frac{\|\hat{\boldsymbol{\Phi}}_{1:k}^l\|_F^2}{\sigma_k^2(\hat{\boldsymbol{\Phi}}_{k,k})} I_G \right] \\ &\leq \exp(-cn) + 2\mathbb{E} \left[\|\hat{\boldsymbol{\Phi}}_{1:k}^l\|^2 I_G \right]. \end{aligned}$$

Now we plug in the upper bounds of $\|\hat{\boldsymbol{\Phi}}_{1:k}^l\|$ obtained in (4.18) and (4.25) respectively. On event G , (4.18) can be reduced to

$$\|\hat{\boldsymbol{\Phi}}_{1:k}^l\|^2 \leq \frac{C}{\Delta^2} (\|\mathbf{B}_1\|^2 + \|\mathbf{B}_2\|^2) + C \|\mathbf{A}_1 \circ \mathbf{A}_2 \circ \boldsymbol{\delta}_6\|^2.$$

Further, by Lemma 4.5 and Lemma 4.6, on event G ,

$$\|\hat{\boldsymbol{\Phi}}_{1:k}^l\|^2 \leq \frac{C}{\Delta^2} (\|\mathbf{B}_1\|^2 + \|\mathbf{B}_2\|^2) + \frac{C}{\Delta^4} \|\boldsymbol{\Sigma}_z - \hat{\boldsymbol{\Sigma}}_z\|^4.$$

Therefore,

$$\mathbb{E} \left[\|\hat{\boldsymbol{\Phi}}_{1:k}^l\|^2 I_G \right] \leq \frac{C}{\Delta^2} \mathbb{E} (\|\mathbf{B}_1\|^2 + \|\mathbf{B}_2\|^2) + \frac{C}{\Delta^4} \mathbb{E} \|\boldsymbol{\Sigma}_z - \hat{\boldsymbol{\Sigma}}_z\|^4.$$

By Lemma 4.2 and Lemma 5.11 (notice that $\|\boldsymbol{\Sigma}_z\| \leq 2$),

$$\mathbb{E} \left[\|\hat{\boldsymbol{\Phi}}_{1:k}^l\|^2 I_G \right] \leq \frac{C(1 - \lambda_k^2)p_1}{\Delta^2 n} + C \left(\frac{p_1 + p_2}{n\Delta^2} \right)^2. \quad (4.31)$$

This upper bound implies the result in Theorem 3.2 when λ_{k+1} is bounded away from 1, for instance, when $\lambda_{k+1} \leq 1/2$. Now, we use (4.25) for the case when $\lambda_{k+1} \geq 1/2$. On event G , with $\lambda_{k+1} \geq 1/2$, (4.25) can be reduced to

$$\|\hat{\Phi}_{1:k}^l\|^2 \leq C \left(\frac{1}{\Delta^2} \|\mathbf{B}^{(1)}\|^2 + \|\mathbf{D}^{-1} \mathbf{B}^{(2)}\|^2 + \|\mathbf{A}_2 \circ \delta_7\|^2 \right).$$

Further apply Lemma 4.5 and Lemma 4.6, on event G ,

$$\|\hat{\Phi}_{1:k}^l\|^2 \leq C \left(\frac{1}{\Delta^2} \|\mathbf{B}^{(1)}\|^2 + \|\mathbf{D}^{-1} \mathbf{B}^{(2)}\|^2 + \frac{1}{\Delta^4} \|\Sigma_z - \hat{\Sigma}_z\|^4 \right).$$

By Lemma 4.3, Lemma 4.4 and Lemma 4.5,

$$\mathbb{E} \left[\|\hat{\Phi}_{1:k}^l\|^2 I_G \right] \leq \frac{C(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)p_1}{\Delta^2 n} + C \left(\frac{p_1 + p_2}{n\Delta^2} \right)^2. \quad (4.32)$$

Combine the results of (4.31) and (4.32),

$$\begin{aligned} \mathbb{E} \left\| \mathbf{P}_{\hat{\Phi}_{1:k}} - \mathbf{P}_{\Phi_{1:k}} \right\|^2 &\leq \exp(-cn) + \frac{C(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)p_1}{\Delta^2 n} + C \left(\frac{p_1 + p_2}{n\Delta^2} \right)^2 \\ &\leq C \frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)p_1}{\Delta^2 n} + C_2 \left(\frac{p_1 + p_2}{n\Delta^2} \right)^2 \end{aligned}$$

4.5 Upper Bound for Frobenius Norm

A quick upper bound in terms of Frobenius norm can be obtained by noticing that $\mathbf{P}_{\hat{\Phi}_{1:k}} - \mathbf{P}_{\Phi_{1:k}}$ has rank at most $2k$ and therefore,

$$\begin{aligned} \frac{1}{k} \mathbb{E} \left\| \mathbf{P}_{\hat{\Phi}_{1:k}} - \mathbf{P}_{\Phi_{1:k}} \right\|_F^2 &\leq 2 \mathbb{E} \left\| \mathbf{P}_{\hat{\Phi}_{1:k}} - \mathbf{P}_{\Phi_{1:k}} \right\|^2 \\ &\leq C \frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)p_1}{\Delta^2 n} + C \left(\frac{p_1 + p_2}{n\Delta^2} \right)^2. \end{aligned}$$

In fact, the factor p_1 in the main term can be reduced to $p_1 - k$ by similar (but much simpler) arguments as done for the operator norm. We state the corresponding results in this section without proof. Specifically, the following Frobenius norm counterparts for (4.18) and (4.25) can be obtained.

Lemma 4.7 *Let $\tilde{\mathbf{D}} = \text{diag}(\lambda_k - \lambda_{k+1}, \dots, \lambda_k - \lambda_{p_1}) \in \mathbb{R}^{(p_1-k) \times (p_1-k)}$, then*

$$\left\| \hat{\Phi}_{1:k}^l \right\|_F \leq \frac{1}{2\Delta} \left(3 \|\mathbf{B}_1\|_F \left\| \hat{\Psi}_{1:k}^u \right\| + \|\mathbf{B}_2\|_F \left\| \hat{\Phi}_{1:k}^u \right\| \right) + \|\mathbf{A}_1 \circ \mathbf{A}_2 \circ \delta_6\|_F, \quad (4.33)$$

$$\left\| \hat{\Phi}_{1:k}^l \right\|_F \leq \frac{1}{\lambda_k} \|\tilde{\mathbf{D}}^{-1} \mathbf{B}\|_F \left\| \hat{\Phi}_{1:k}^u \right\| + \frac{1}{\lambda_k} \|\mathbf{A}_2 \circ \delta_7\|_F. \quad (4.34)$$

Notice that for the second bound, divide-and-conquer type of analysis in the proof for operator norm is no longer necessary due to the observation that

$$\|\mathbf{A} \circ \mathbf{M}\|_F \leq \|\mathbf{M}\|_F,$$

when $\max_{i,j} |\mathbf{A}_{ij}| \leq 1$ while the inequality is not true for operator norm. Similarly, parallel results to the lemmas in Section 4.3 can be derived as follows (see Section 5.7 for the proof of the second inequality in the lemma as illustration).

Lemma 4.8 *There exists universal constant C such that*

$$\begin{aligned} \mathbb{E}[\|\mathbf{B}_1\|_F^2], \mathbb{E}[\|\mathbf{B}_2\|_F^2] &\leq C \frac{(1 - \lambda_k^2)(p_1 - k)k}{n}, \\ \mathbb{E}\left[\left\|\tilde{\mathbf{D}}^{-1}\mathbf{B}\right\|_F^2\right] &\leq 2 \frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)(p_1 - k)k}{n\Delta^2}, \\ \|\mathbf{A}_1 \circ \mathbf{A}_2 \circ \delta_6\|_F, \|\mathbf{A}_2 \circ \delta_7\|_F &\leq \frac{C\sqrt{k}\|\boldsymbol{\Sigma}_z - \hat{\boldsymbol{\Sigma}}_z\|^2}{\Delta^2} (2 + \|\hat{\boldsymbol{\Sigma}}_z\|)^2 (\|\hat{\boldsymbol{\Phi}}_{1:k}\| + \|\hat{\boldsymbol{\Psi}}_{1:k}\|) (1 + \|\boldsymbol{\Sigma}_z - \hat{\boldsymbol{\Sigma}}_z\|). \end{aligned}$$

Substituting these lemmas into the procedure of Section 4.4 will obtain

$$\frac{1}{k} \mathbb{E} \left\| \mathbf{P}_{\hat{\boldsymbol{\Phi}}_{1:k}} - \mathbf{P}_{\boldsymbol{\Phi}_{1:k}} \right\|_F^2 \leq \frac{C(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)(p_1 - k)}{\Delta^2 n} + C \left(\frac{p_1 + p_2}{n\Delta^2} \right)^2.$$

5 Proof of the Main Results

5.1 Proof of Theorem 2.1

For any $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{R}^{p \times k}$, from classical linear model theory,

$$\boldsymbol{\beta}_i := \arg \min_{\boldsymbol{\beta}} \mathbb{E} \left[\left| z - \boldsymbol{\beta}^\top (\mathbf{U}_i^\top \mathbf{x}_i) \right|^2 \right] = (\mathbf{U}_i^\top \boldsymbol{\Sigma}_x \mathbf{U}_i)^{-1} \mathbf{U}_i^\top \boldsymbol{\Sigma}_{xz}, \quad i = 1, 2,$$

and

$$\begin{aligned} \text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U}_i)) &= \mathbb{E}[z^2] - \mathbb{E}[(\boldsymbol{\beta}_i^\top (\mathbf{U}_i^\top \mathbf{x}_i))^2] \\ &= \sigma_z^2 - \boldsymbol{\Sigma}_{zx} \mathbf{U}_i (\mathbf{U}_i^\top \boldsymbol{\Sigma}_x \mathbf{U}_i)^{-1} \mathbf{U}_i^\top \boldsymbol{\Sigma}_x \mathbf{U}_i (\mathbf{U}_i^\top \boldsymbol{\Sigma}_x \mathbf{U}_i)^{-1} \mathbf{U}_i^\top \boldsymbol{\Sigma}_{xz} \\ &= \sigma_z^2 \left(1 - \mathbf{r}_{zx} (\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_i) (\mathbf{U}_i^\top \boldsymbol{\Sigma}_x \mathbf{U}_i)^{-1} (\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_i)^\top \mathbf{r}_{xz} \right). \end{aligned}$$

Notice that $(\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_i) (\mathbf{U}_i^\top \boldsymbol{\Sigma}_x \mathbf{U}_i)^{-1/2}$ has orthonormal columns with column space $\text{span}(\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_i)$, then

$$\text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U}_i)) = \sigma_z^2 \left(1 - \mathbf{r}_{zx} \mathbf{P}_{\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_i} \mathbf{r}_{xz} \right).$$

Therefore,

$$\text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U}_1)) - \text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U}_2)) = \sigma_z^2 \mathbf{r}_{zx} \left(\mathbf{P}_{\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_2} - \mathbf{P}_{\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_1} \right) \mathbf{r}_{xz}.$$

By the variational definition of the leading eigenvalue,

$$\begin{aligned} \sup_{\|\mathbf{r}_{xz}\|^2=R^2} \text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U}_1)) - \text{loss}(z | \text{span}(\mathbf{x}^\top \mathbf{U}_2)) &= \sigma_z^2 R^2 \lambda_{\max}(\mathbf{P}_{\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_2} - \mathbf{P}_{\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_1}) \\ &= \sigma_z^2 R^2 \left\| \mathbf{P}_{\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_2} - \mathbf{P}_{\boldsymbol{\Sigma}_x^{1/2} \mathbf{U}_1} \right\|, \end{aligned}$$

where the second inequality is by the characterization of the difference between two projection matrices described in [Wedin \[1983\]](#). This proves the first claim of the theorem. For the second part, since $\mathbf{r}_{xz} \in \mathcal{A}$,

$$\begin{aligned}
& \text{loss}(z|\text{span}(\mathbf{x}^\top \mathbf{U})) - \text{loss}(z|\text{span}(\mathbf{x}^\top \mathbf{U}_*)) \\
&= \sigma_z^2 \mathbf{r}_{xz}^\top \left(\mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} \right) \mathbf{r}_{xz} \\
&= \sigma_z^2 \mathbf{r}_{xz}^\top \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \left(\mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} \right) \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \mathbf{r}_{xz} \\
&= \sigma_z^2 \mathbf{r}_{xz}^\top \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \left(\mathbf{I}_p - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} \right) \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \mathbf{r}_{xz} \\
&\leq \sigma_z^2 R^2 \left\| \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \left(\mathbf{I}_p - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} \right) \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \right\| \\
&= \sigma_z^2 R^2 \left\| \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \left(\mathbf{I}_p - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} \right) \right\|^2 \\
&= \sigma_z^2 R^2 \left\| \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} \right\|^2.
\end{aligned}$$

Notice that $\mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \left(\mathbf{I}_p - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} \right) \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*}$ is positive definite. Let \mathbf{u}_1 be the leading singular vector of this matrix and define $\mathbf{r}_{xz}^* = R \mathbf{u}_1$. Then $\mathbf{r}_{xz}^* \in \mathcal{A}$ and with such choice of \mathbf{r}_{xz}^* , the inequality above will become equality and this implies that

$$\sup_{\mathbf{r}_{xz}^* \in \mathcal{A}} \left\{ \text{loss}(z|\text{span}(\mathbf{x}^\top \mathbf{U})) - \text{loss}(z|\text{span}(\mathbf{x}^\top \mathbf{U}_*)) \right\} = \sigma_z^2 R^2 \left\| \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \right\|^2.$$

For the last part of the theorem, let $\mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} = \mathbf{Q} \mathbf{Q}^\top$. Then the set \mathcal{A} can be rewritten as

$$\mathcal{A} = \{ \mathbf{r} : \mathbf{r} = R^2 \mathbf{Q} \tilde{\mathbf{r}}, \tilde{\mathbf{r}} \in \mathbb{S}^{k-1} \}.$$

The uniform measure (Haar measure) π on \mathcal{A} can be expressed through the uniform measure (Haar measure) $\tilde{\pi}$ on the sphere \mathbb{S}^{k-1} . Because $\tilde{\pi}$ is uniform on the sphere \mathbb{S}^{k-1} , then $\text{Var}(\tilde{\pi}) = \mathbf{I}_k/k$ and

$$\begin{aligned}
& \mathbb{E}_{\mathbf{r}_{xz} \sim \pi} [\text{loss}(z|\text{span}(\mathbf{x}^\top \mathbf{U})) - \text{loss}(z|\text{span}(\mathbf{x}^\top \mathbf{U}_*))] \\
&= \sigma_z^2 R^2 \mathbb{E}_{\tilde{\mathbf{r}} \sim \tilde{\pi}} \left[\tilde{\mathbf{r}}^\top \mathbf{Q}^\top \left(\mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} \right) \mathbf{Q} \tilde{\mathbf{r}} \right] \\
&= \sigma_z^2 R^2 \text{tr} \left(\mathbf{Q}^\top \left(\mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} \right) \mathbf{Q} \right) / k \\
&= \sigma_z^2 R^2 \text{tr} \left(\left(\mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} \right) \mathbf{Q} \mathbf{Q}^\top \right) / k \\
&= \sigma_z^2 R^2 \text{tr} \left(\left(\mathbf{I}_p - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} \right) \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \right) / k \\
&= \sigma_z^2 R^2 \text{tr} \left(\mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \left(\mathbf{I}_p - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} \right) \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \right) / k \\
&= \sigma_z^2 R^2 \left\| \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \left(\mathbf{I}_p - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} \right) \right\|_F^2 / k \\
&= \sigma_z^2 R^2 \left\| \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}} - \mathbf{P}_{\Sigma_x^{1/2} \mathbf{U}_*} \right\|_F^2 / 2k,
\end{aligned}$$

where the last equality is due to Lemma 5.8.

5.2 Proof of Lemma 4.1

The proof of Lemma 4.1 relies on the following two results.

Lemma 5.1 (*Hom and Johnson [1991], Theorem 5.5.18*) If $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ and \mathbf{A} is positive semidefinite. Then,

$$\|\mathbf{A} \circ \mathbf{B}\| \leq \left(\max_{1 \leq i \leq n} A_{ii} \right) \|\mathbf{B}\|,$$

where $\|\cdot\|$ is the operator norm.

Lemma 5.2 (*Fiedler [2010], Theorem A*) A symmetric Cauchy matrix

$$\mathbf{C} = \left(\frac{1}{a_i + a_j} \right)_{1 \leq i, j \leq n}$$

is positive semidefinite if $a_i > 0, 1 \leq i \leq n$.

Define $\gamma_i = \beta_i, 1 \leq i \leq n$ and $\gamma_i = \alpha_{i-n}, n+1 \leq i \leq m+n$. Consider the matrix $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{(m+n) \times (m+n)}$ define by

$$[\mathbf{M}_1]_{ij} = \frac{1}{\gamma_i + \gamma_j}, \quad [\mathbf{M}_2]_{ij} = \frac{\min\{\gamma_i, \gamma_j\}}{\gamma_i + \gamma_j}.$$

By Lemma 5.2, \mathbf{M}_1 is positive semidefinite and by Lemma 5.1,

$$\|\mathbf{M}_1\| \leq \frac{1}{2 \min_{1 \leq i \leq m+n} \{\gamma_i\}} \leq \frac{1}{2\delta}.$$

Notice that \mathbf{A}_1 is the lower left sub-matrix of \mathbf{M}_1 , therefore,

$$\|\mathbf{A}_1\| \leq \|\mathbf{M}_1\| \leq \frac{1}{2\delta}.$$

By Theorem 3.2 of Mathias [1993], \mathbf{M}_2 is also positive semidefinite. Again, apply Lemma 5.1 and notice that \mathbf{A}_2 is the lower left sub-matrix of \mathbf{M}_2 ,

$$\|\mathbf{A}_2\| \leq \|\mathbf{M}_2\| \leq \frac{1}{2}.$$

Finally, observe that, by definition, $\mathbf{A}_3 \circ \mathbf{B} = \mathbf{B} - \mathbf{A}_2 \circ \mathbf{B}$, hence

$$\|\mathbf{A}_3 \circ \mathbf{B}\| \leq \|\mathbf{B}\| + \|\mathbf{A}_2 \circ \mathbf{B}\|,$$

which implies,

$$\|\mathbf{A}_3\| \leq 1 + \|\mathbf{A}_2\| \leq \frac{3}{2}.$$

5.3 Proof of Lemma 4.2

Divide \mathbf{x}, \mathbf{y} into two parts,

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix},$$

where $\mathbf{x}_1, \mathbf{y}_1 \in \mathbb{R}^k$, $\mathbf{x}_2 \in \mathbb{R}^{p_1-k}$ and $\mathbf{y}_2 \in \mathbb{R}^{p_2-k}$. By definition,

$$\mathbf{B}_1 = \widehat{\text{Cov}}(\mathbf{x}_2, \mathbf{y}_1 - \mathbf{\Lambda}_1 \mathbf{x}_1),$$

where $\widehat{\text{Cov}}(\cdot, \cdot)$ denotes the sample covariance operator. When $\lambda_k = 1$, by definition of CCA, $\mathbf{y}_1 = \mathbf{x}_1$ almost surely, which implies that $\mathbf{B}_1 = \mathbf{0}$ almost surely. When $\lambda_k < 1$,

$$\mathbf{B}_1 = \widehat{\text{Cov}}(\mathbf{x}_2, \mathbf{y}_1 - \mathbf{\Lambda}_1 \mathbf{x}_1) = \sqrt{1 - \lambda_k^2} \widehat{\text{Cov}} \left(\mathbf{x}_2, \frac{\mathbf{y}_1 - \mathbf{\Lambda}_1 \mathbf{x}_1}{\sqrt{1 - \lambda_k^2}} \right) = \sqrt{1 - \lambda_k^2} \widehat{\text{Cov}}(\mathbf{w}_1, \mathbf{w}_2),$$

where we define $\mathbf{w}_1 = \frac{\mathbf{y}_1 - \mathbf{\Lambda}_1 \mathbf{x}_1}{\sqrt{1 - \lambda_k^2}}$ and $\mathbf{w}_2 = \mathbf{x}_2$. Let $\mathbf{w} = (\mathbf{w}_1^\top, \mathbf{w}_2^\top)^\top$ be the concatenation of \mathbf{w}_1 and \mathbf{w}_2 . Then

$$\mathbf{\Sigma}_w = \text{Var}(w) = \text{diag} \left(\frac{1 - \lambda_1^2}{1 - \lambda_k^2}, \frac{1 - \lambda_2^2}{1 - \lambda_k^2}, \dots, \frac{1 - \lambda_k^2}{1 - \lambda_k^2}, 1, \dots, 1 \right).$$

Notice that $\|\mathbf{\Sigma}_w\| \leq 1$ and $\mathbb{E}[\widehat{\text{Cov}}(\mathbf{w}_1, \mathbf{w}_2)] = \mathbf{0}$. Therefore,

$$\|\mathbf{B}_1\|^2 \leq (1 - \lambda_k^2) \|\widehat{\text{Var}}(\mathbf{w}) - \mathbf{\Sigma}_w\|^2.$$

By Lemma 5.11,

$$\mathbb{E}[\|\mathbf{B}_1\|^2] \leq C \frac{(1 - \lambda_k^2)p_1}{n}.$$

Similarly, we can prove the result for \mathbf{B}_2 in the same manner.

5.4 Proof of Lemma 4.3

Step 1. Reduction. For $\epsilon > 0$ and any pair of vectors $\mathbf{u} \in \mathbb{R}^{p_1-k^*}$, $\mathbf{v} \in \mathbb{R}^k$, we can choose $\mathbf{u}_\epsilon \in \mathcal{N}(\mathbb{S}^{p_1-k^*-1}, \epsilon)$, $\mathbf{v}_\epsilon \in \mathcal{N}(\mathbb{S}^{k-1}, \epsilon)$ such that $\|\mathbf{u} - \mathbf{u}_\epsilon\|, \|\mathbf{v} - \mathbf{v}_\epsilon\| \leq \epsilon$. Then

$$\begin{aligned} \mathbf{u}^\top \mathbf{D}^{-1} \mathbf{B}^{(2)} \mathbf{v} &= \mathbf{u}^\top \mathbf{D}^{-1} \mathbf{B}^{(2)} \mathbf{v} - \mathbf{u}_\epsilon^\top \mathbf{D}^{-1} \mathbf{B}^{(2)} \mathbf{v} + \mathbf{u}_\epsilon^\top \mathbf{D}^{-1} \mathbf{B}^{(2)} \mathbf{v} - \mathbf{u}_\epsilon^\top \mathbf{D}^{-1} \mathbf{B}^{(2)} \mathbf{v}_\epsilon + \mathbf{u}_\epsilon^\top \mathbf{D}^{-1} \mathbf{B}^{(2)} \mathbf{v}_\epsilon \\ &\leq \|\mathbf{u} - \mathbf{u}_\epsilon\| \|\mathbf{D}^{-1} \mathbf{B}^{(2)} \mathbf{v}\| + \|\mathbf{u}_\epsilon^\top \mathbf{D}^{-1} \mathbf{B}^{(2)}\| \|\mathbf{v} - \mathbf{v}_\epsilon\| + \mathbf{u}_\epsilon^\top \mathbf{D}^{-1} \mathbf{B}^{(2)} \mathbf{v}_\epsilon \\ &\leq 2\epsilon \|\mathbf{D}^{-1} \mathbf{B}^{(2)}\| + \mathbf{u}_\epsilon^\top \mathbf{D}^{-1} \mathbf{B}^{(2)} \mathbf{v}_\epsilon \\ &\leq 2\epsilon \|\mathbf{D}^{-1} \mathbf{B}^{(2)}\| + \max_{\mathbf{u}_\epsilon, \mathbf{v}_\epsilon} \mathbf{u}_\epsilon^\top \mathbf{D}^{-1} \mathbf{B}^{(2)} \mathbf{v}_\epsilon. \end{aligned}$$

Maximize over \mathbf{u} and \mathbf{v} , we obtain

$$\|\mathbf{D}^{-1} \mathbf{B}^{(2)}\| \leq 2\epsilon \|\mathbf{D}^{-1} \mathbf{B}^{(2)}\| + \max_{\mathbf{u}_\epsilon, \mathbf{v}_\epsilon} \mathbf{u}_\epsilon^\top \mathbf{D}^{-1} \mathbf{B}^{(2)} \mathbf{v}_\epsilon. \quad (5.1)$$

Therefore, $\|D^{-1}B^{(2)}\| \leq (1 - 2\epsilon)^{-1} \max_{\mathbf{u}_\epsilon, \mathbf{v}_\epsilon} \mathbf{u}_\epsilon^\top D^{-1}B^{(2)}\mathbf{v}_\epsilon$. Let $\epsilon = 1/4$. Then it suffices to prove with required probability,

$$\max_{\mathbf{u}_\epsilon, \mathbf{v}_\epsilon} \mathbf{u}_\epsilon^\top D^{-1}B^{(2)}\mathbf{v}_\epsilon \leq \frac{1}{2} \max\{\delta, \delta^2\}. \quad (5.2)$$

Step 2. Concentration. Notice that for $1 \leq j \leq k < k^* + 1 \leq i \leq p_1$,

$$\begin{aligned} [D^{-1}B^{(2)}]_{i-k^*,j} &= \frac{1}{\lambda_k - \Delta/2 - \lambda_i} \frac{1}{n} \sum_{\alpha=1}^n (\lambda_j \mathbf{x}_{\alpha i} \mathbf{y}_{\alpha j} - \lambda_j^2 \mathbf{x}_{\alpha i} \mathbf{x}_{\alpha j} + \lambda_i \mathbf{x}_{\alpha j} \mathbf{y}_{\alpha i} - \lambda_i \lambda_j \mathbf{y}_{\alpha i} \mathbf{y}_{\alpha j}) \\ &= \frac{1}{\lambda_k - \Delta/2 - \lambda_i} \frac{1}{n} \sum_{\alpha=1}^n \left\{ (1 - \lambda_j^2) \lambda_i \lambda_j \mathbf{x}_{\alpha i} \mathbf{x}_{\alpha j} - \lambda_j^2 (\mathbf{y}_{\alpha i} - \lambda_i \mathbf{x}_{\alpha i}) (\mathbf{y}_{\alpha j} - \lambda_j \mathbf{x}_{\alpha j}) \right. \\ &\quad \left. + (1 - \lambda_j^2) \lambda_j (\mathbf{y}_{\alpha i} - \lambda_i \mathbf{x}_{\alpha i}) \mathbf{x}_{\alpha j} + (1 - \lambda_j^2) \lambda_i (\mathbf{y}_{\alpha j} - \lambda_j \mathbf{x}_{\alpha j}) \mathbf{x}_{\alpha i} \right\}. \end{aligned}$$

Let $\mathbf{z}_l = (\mathbf{y}_l - \lambda_i \mathbf{x}_l) / \sqrt{1 - \lambda_i^2}$, $1 \leq l \leq p_1$. Then

$$\begin{aligned} [D^{-1}B^{(2)}]_{i-k^*,j} &= \frac{1}{\lambda_k - \Delta/2 - \lambda_i} \frac{1}{n} \sum_{\alpha=1}^n \left\{ (1 - \lambda_j^2) \lambda_i \lambda_j \mathbf{x}_{\alpha i} \mathbf{x}_{\alpha j} - \lambda_j^2 \sqrt{1 - \lambda_i^2} \sqrt{1 - \lambda_j^2} \mathbf{z}_{\alpha i} \mathbf{z}_{\alpha j} \right. \\ &\quad \left. + (1 - \lambda_j^2) \lambda_j \sqrt{1 - \lambda_i^2} \mathbf{z}_{\alpha i} \mathbf{x}_{\alpha j} + (1 - \lambda_j^2) \lambda_i \sqrt{1 - \lambda_j^2} \mathbf{z}_{\alpha j} \mathbf{x}_{\alpha i} \right\}. \end{aligned}$$

In this way, $\{\mathbf{x}_{\alpha i}, \mathbf{z}_{\alpha i}, 1 \leq i \leq p_1, 1 \leq \alpha \leq n\}$ are mutually independent standard gaussian random variables. For any given pair of vectors $\mathbf{u} \in \mathbb{R}^{p_1-k^*}$, $\mathbf{v} \in \mathbb{R}^k$,

$$\begin{aligned} \mathbf{u}^\top D^{-1}B^{(2)}\mathbf{v} &= \frac{1}{n} \sum_{\alpha=1}^n \sum_{i=k^*+1}^{p_1} \sum_{j=1}^k \frac{\mathbf{u}_{i-k} \mathbf{v}_j}{\lambda_k - \Delta/2 - \lambda_i} \left\{ (1 - \lambda_j^2) \lambda_i \lambda_j \mathbf{x}_{\alpha i} \mathbf{x}_{\alpha j} - \lambda_j^2 \sqrt{1 - \lambda_i^2} \sqrt{1 - \lambda_j^2} \mathbf{z}_{\alpha i} \mathbf{z}_{\alpha j} \right. \\ &\quad \left. + (1 - \lambda_j^2) \lambda_j \sqrt{1 - \lambda_i^2} \mathbf{z}_{\alpha i} \mathbf{x}_{\alpha j} + (1 - \lambda_j^2) \lambda_i \sqrt{1 - \lambda_j^2} \mathbf{z}_{\alpha j} \mathbf{x}_{\alpha i} \right\} \\ &\doteq \frac{1}{n} \sum_{\alpha=1}^n \mathbf{w}_\alpha, \end{aligned}$$

where $\mathbf{w}_1, \dots, \mathbf{w}_n$ are *i.i.d.* quadratic forms of the concatenated vector $(\mathbf{x}^\top, \mathbf{z}^\top)$ and the quadratic form can be represented by a matrix \mathbf{A} . In order to apply Lemma 5.13, we first compute,

$$\begin{aligned} \|\mathbf{A}\|_F^2 &= \sum_{i=k^*+1}^{p_1} \sum_{j=1}^k \frac{\mathbf{u}_{i-k}^2 \mathbf{v}_j^2}{(\lambda_k - \Delta/2 - \lambda_i)^2} \left\{ (1 - \lambda_j^2)^2 \lambda_i^2 \lambda_j^2 + \lambda_j^4 (1 - \lambda_i^2) (1 - \lambda_j^2) \right. \\ &\quad \left. + (1 - \lambda_j^2)^2 \lambda_j^2 (1 - \lambda_i^2) + (1 - \lambda_j^2)^2 \lambda_i^2 (1 - \lambda_j^2) \right\} \\ &= \sum_{i=k^*+1}^{p_1} \sum_{j=1}^k \frac{\mathbf{u}_{i-k}^2 \mathbf{v}_j^2}{(\lambda_k - \Delta/2 - \lambda_i)^2} (1 - \lambda_j^2) (\lambda_i^2 + \lambda_j^2 - 2\lambda_i^2 \lambda_j^2). \end{aligned}$$

By definition,

$$\sum_{i=k^*+1}^{p_1} \sum_{j=1}^k \mathbf{u}_{i-k}^2 \mathbf{v}_j^2 = 1.$$

Then $\|\mathbf{A}\|_2^2, \|\mathbf{A}\|_F^2$ can be upper bounded by

$$\begin{aligned}
\|\mathbf{A}\|_2^2 &\leq \|\mathbf{A}\|_F^2 \leq \max_{1 \leq j \leq k < k^*+1 \leq i \leq p_1} \frac{(1 - \lambda_j^2)(\lambda_i^2 + \lambda_j^2 - 2\lambda_i^2\lambda_j^2)}{(\lambda_k - \Delta/2 - \lambda_i)^2} \\
&\leq \max_{1 \leq j \leq k < k^*+1 \leq i \leq p_1} \frac{(1 - \lambda_k^2)(\lambda_i^2(1 - \lambda_j^2) + \lambda_j^2(1 - \lambda_i^2))}{(\lambda_k - \Delta/2 - \lambda_i)^2} \\
&\leq (1 - \lambda_k^2) \max_{1 \leq j \leq k < k^*+1 \leq i \leq p_1} \frac{2(1 - \lambda_i^2)}{(\lambda_k - \Delta/2 - \lambda_i)^2} \\
&\leq (1 - \lambda_k^2) \max_{1 \leq j \leq k < k^*+1 \leq i \leq p_1} \frac{2(1 - \lambda_{k+1}^2)}{(\lambda_k - \Delta/2 - \lambda_{k+1})^2} \\
&\leq 8 \frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{\Delta^2} \doteq K^2,
\end{aligned}$$

where the second last inequality is due to the fact that for $\lambda > \lambda_{k+1}$, $f(x) = \frac{1-x^2}{(\lambda-x)^2}$ is increasing in the interval $[0, \lambda_{k+1}]$. Therefore, Lemma 5.13 implies that

$$P\{|\mathbf{w}_\alpha| \geq t\} \leq 2\exp\left\{-c_0 \min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right\}. \quad (5.3)$$

Observe that $\forall t \geq 0, \min\left(1, 2\exp\left\{-c_0 \min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right\}\right) \leq \exp\{1 - c_0 \frac{t}{K}\}$, then

$$P\{|\mathbf{w}_\alpha| \geq t\} \leq 2\exp\left\{-c_0 \left(\frac{t}{K} - 1\right)\right\}. \quad (5.4)$$

By Definition 5.13 in Vershynin [2010], \mathbf{w}_α is sub-exponential random variable with $\|\mathbf{w}_\alpha\|_{\psi_1} \leq c_1 K$ for some universal constant c_1 . Let $\tilde{\delta} = \max\{\delta, \delta^2\}$. Apply Bernstein inequality (Lemma 5.12) to \mathbf{w}_α/K with $\mathbf{a}_i = 1/\sqrt{n}$,

$$\begin{aligned}
P\left\{\left|\frac{1}{\sqrt{n}} \sum_{\alpha=1}^n \mathbf{w}_\alpha/K\right| \geq \tilde{\delta}/2\right\} &\leq 2\exp\left\{-c_2 n \min\left(\frac{\tilde{\delta}^2}{4c_1^2}, \frac{\tilde{\delta}}{2c_1}\right)\right\} \\
&\leq 2\exp\left\{-\frac{c_2}{1+4c_1^2} n \delta^2\right\} \\
&\leq 2\exp\left\{-\frac{c_2}{1+4c_1^2} (C^2 p_1 + t^2)\right\}.
\end{aligned}$$

Step 3. Union Bound. By Lemma 5.14, we can choose 1/4-net such that

$$\begin{aligned}
P\left\{\max_{\mathbf{u}_\epsilon, \mathbf{v}_\epsilon} \mathbf{u}_\epsilon^\top \mathbf{B}^{(2)} \mathbf{v}_\epsilon \geq K\tilde{\delta}/2\right\} &\leq 9^{p_1-k^*} 9^k \times 2\exp\left\{-\frac{c_2}{1+4c_1^2} (C^2 p_1 + t^2)\right\} \\
&\leq 2\exp\left\{-\frac{c_2}{1+4c_1^2} t^2\right\},
\end{aligned}$$

where the second inequality follows if we choose $C \geq \sqrt{\frac{(1+4c_1^2)\log 9}{c_2}}$. We finish the proof by choosing $c = \frac{c_2}{1+4c_1^2}$. The expectation bound can be obtained using the formula

$$E[X] = \int_0^{+\infty} P(X \geq t) dt$$

where X is nonnegative random variable. The calculation is essentially the same as the proof in Section 6.5 and we leave out the details.

5.5 Proof of Lemma 4.4

The proof is essentially the same as the proof for Lemma 4.3 except that the term $\|\mathbf{A}\|_F^2$ will be different (but simpler) and is sketched as follows,

$$\begin{aligned}
\|\mathbf{A}\|_F^2 &= \sum_{i=k+1}^{k^*} \sum_{j=1}^k \mathbf{u}_{i-k}^2 \mathbf{v}_j^2 \left\{ (1 - \lambda_j^2)^2 \lambda_i^2 \lambda_j^2 + \lambda_j^4 (1 - \lambda_i^2) (1 - \lambda_j^2) \right. \\
&\quad \left. + (1 - \lambda_j^2)^2 \lambda_j^2 (1 - \lambda_i^2) + (1 - \lambda_j^2)^2 \lambda_i^2 (1 - \lambda_i^2) \right\} \\
&= \sum_{i=k+1}^{k^*} \sum_{j=1}^k \mathbf{u}_{i-k}^2 \mathbf{v}_j^2 (1 - \lambda_j^2) (\lambda_i^2 + \lambda_j^2 - 2\lambda_i^2 \lambda_j^2) \\
&\leq \max_{1 \leq j \leq k < i \leq k^*} (1 - \lambda_k^2) (\lambda_i^2 (1 - \lambda_j^2) + \lambda_j^2 (1 - \lambda_i^2)) \\
&\leq 2(1 - \lambda_k^2) \max_{1 \leq j \leq k < i \leq k^*} (1 - \lambda_i^2).
\end{aligned}$$

Notice that by definition, for $k+1 \leq i \leq k^*$, $\lambda_i \geq 2\lambda_k - 1 - \Delta$, then

$$\begin{aligned}
\|\mathbf{A}\|_F^2 &\leq 2(1 - \lambda_k^2) (1 + \lambda_i) (2 - 2\lambda_k + \Delta) \\
&\leq 2(1 - \lambda_k^2) (1 + \lambda_{k+1}) 2(1 - \lambda_{k+1}) \\
&\leq 4(1 - \lambda_k^2) (1 - \lambda_{k+1}^2).
\end{aligned}$$

The rest of the argument proceeds in the same way as in the proof of Lemma 4.3.

5.6 Proof of Lemma 4.6

By definition, $\hat{\mathbf{\Phi}}^\top \hat{\Sigma}_x \hat{\mathbf{\Phi}} = \mathbf{I}_{p_1}$, then

$$\hat{\mathbf{\Phi}}^\top \hat{\mathbf{\Phi}} - \mathbf{I}_{p_1} = -\hat{\mathbf{\Phi}}^\top (\hat{\Sigma}_x - \mathbf{I}_{p_1}) \hat{\mathbf{\Phi}}.$$

Notice that $\hat{\Sigma}_x^{1/2} \hat{\mathbf{\Phi}} \in \mathcal{O}(p_1)$,

$$\begin{aligned}
\|\hat{\mathbf{\Phi}}^\top \hat{\mathbf{\Phi}} - \mathbf{I}_{p_1}\| &\leq \|\hat{\mathbf{\Phi}}^\top (\hat{\Sigma}_x - \mathbf{I}_{p_1}) \hat{\mathbf{\Phi}}\| \leq \|\hat{\mathbf{\Phi}}^\top \hat{\Sigma}_x^{1/2}\| \|\hat{\Sigma}_x^{-1/2} (\hat{\Sigma}_x - \mathbf{I}_{p_1}) \hat{\Sigma}_x^{-1/2}\| \|\hat{\Sigma}_x^{1/2} \hat{\mathbf{\Phi}}\| \\
&= \|\hat{\Sigma}_x^{-1/2} (\hat{\Sigma}_x - \mathbf{I}_{p_1}) \hat{\Sigma}_x^{-1/2}\|.
\end{aligned}$$

As a submatrix,

$$\begin{aligned}
\|\hat{\mathbf{\Phi}}_{1:k}^\top \hat{\mathbf{\Phi}}_{1:k} - \mathbf{I}_k\| &\leq \|\hat{\Sigma}_x^{-1/2} (\hat{\Sigma}_x - \mathbf{I}_{p_1}) \hat{\Sigma}_x^{-1/2}\| \\
&\leq \|\hat{\Sigma}_x^{-1}\| \|\hat{\Sigma}_x - \mathbf{I}_{p_1}\| \\
&\leq \frac{1}{1 - \|\hat{\Sigma}_x - \mathbf{I}_{p_1}\|} \|\hat{\Sigma}_x - \mathbf{I}_{p_1}\| \\
&\leq \frac{\|\hat{\Sigma}_z - \Sigma_z\|}{1 - \|\hat{\Sigma}_z - \Sigma_z\|},
\end{aligned}$$

which implies that

$$\sigma_k^2(\hat{\Phi}_{1:k}) \geq 1 - \frac{\|\hat{\Sigma}_z - \Sigma_z\|}{1 - \|\hat{\Sigma}_z - \Sigma_z\|}, \quad \|\hat{\Phi}_{1:k}\|^2 \leq 1 + \frac{\|\hat{\Sigma}_z - \Sigma_z\|}{1 - \|\hat{\Sigma}_z - \Sigma_z\|}.$$

Notice that $\|\Sigma_z\| \leq 2$. By Lemma 5.10, for any given positive constant τ , there exists constants c, C depending on τ such that when $n \geq C(p_1 + p_2)$,

$$\|\hat{\Sigma}_z - \Sigma_z\| \leq \tau$$

holds with probability at least $1 - e^{-cn}$. Choose τ small enough such that $\sigma_k^2(\hat{\Phi}_{1:k}) \geq 1/2$ and $\|\hat{\Phi}_{1:k}\|^2 \leq 3/2$. By the same argument,

$$\sigma_k^2(\hat{\Psi}_{1:k}) \geq 1/2, \quad \|\hat{\Psi}_{1:k}\|^2 \leq 3/2$$

will hold as well and therefore

$$(2 + \|\hat{\Sigma}_z\|)^2(\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \hat{\Sigma}_z\|) \leq (2 + \|\Sigma_z\| + \tau)^2 \times 3 \times (1 + \tau) \leq 150$$

5.7 Proof of Lemma 4.8

We can write down explicitly, for $1 \leq j \leq k < i \leq p_1$

$$[B]_{i-k,j} = \frac{1}{n} \sum_{\alpha=1}^n (\lambda_j \mathbf{x}_{\alpha i} \mathbf{y}_{\alpha j} - \lambda_j^2 \mathbf{x}_{\alpha i} \mathbf{x}_{\alpha j} + \lambda_i \mathbf{x}_{\alpha j} \mathbf{y}_{\alpha i} - \lambda_i \lambda_j \mathbf{y}_{\alpha i} \mathbf{y}_{\alpha j}).$$

Notice that $(\mathbf{x}_{\alpha i}, \mathbf{y}_{\alpha i})$ are mutually uncorrelated pairs for any $1 \leq \alpha \leq n, 1 \leq i \leq p_1$. It is easy to compute

$$\mathbb{E}[B]_{i-k,j}^2 = (1 - \lambda_j^2)(\lambda_i^2 + \lambda_j^2 - 2\lambda_i^2 \lambda_j^2)/n,$$

and thus

$$\begin{aligned} \mathbb{E}\|\tilde{D}^{-1}B\|_F^2 &= \frac{1}{n} \sum_{1 \leq j \leq k < i \leq p_1} (1 - \lambda_j^2) \frac{\lambda_i^2 + \lambda_j^2 - 2\lambda_i^2 \lambda_j^2}{(\lambda_k - \lambda_i)^2} \\ &\leq \frac{1 - \lambda_k^2}{n} \sum_{1 \leq j \leq k < i \leq p_1} \frac{\lambda_i^2(1 - \lambda_j^2) + \lambda_j^2(1 - \lambda_i^2)}{(\lambda_k - \lambda_i)^2} \\ &\leq \frac{2(1 - \lambda_k^2)}{n} \sum_{1 \leq j \leq k < i \leq p_1} \frac{1 - \lambda_i^2}{(\lambda_k - \lambda_i)^2} \\ &\leq \frac{2(1 - \lambda_k^2)}{n} \sum_{1 \leq j \leq k < i \leq p_1} \frac{1 - \lambda_{k+1}^2}{(\lambda_k - \lambda_{k+1})^2} \\ &\leq \frac{2(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{n\Delta^2} (p_1 - k)k, \end{aligned}$$

where the second last inequality is due to the fact that for $\lambda > \lambda_{k+1}$, $f(x) = \frac{1-x^2}{(\lambda-x)^2}$ is increasing in the interval $[0, \lambda_{k+1}]$.

5.8 Lower Bound: Proof of Theorem 3.3

5.8.1 On Kullback-Leibler Divergence

Lemma 5.3 For $i = 1, 2$ and $p_2 \geq p_1 \geq k$, let $[U_{(i)}, W_{(i)}] \in \mathcal{O}(p_1, p_1)$, $[V_{(i)}, Z_{(i)}] \in \mathcal{O}(p_2, p_1)$ where $U_{(i)} \in \mathbb{R}^{p_1 \times k}$, $V_{(i)} \in \mathbb{R}^{p_2 \times k}$. For $0 \leq \lambda_2 < \lambda_1 < 1$, let $\Delta = \lambda_1 - \lambda_2$ and define

$$\Sigma_{(i)} = \begin{bmatrix} \Sigma_x & \Sigma_x^{1/2}(\lambda_1 U_{(i)} V_{(i)}^\top + \lambda_2 W_{(i)} Z_{(i)}^\top) \Sigma_y^{1/2} \\ \Sigma_y^{1/2}(\lambda_1 V_{(i)} U_{(i)}^\top + \lambda_2 Z_{(i)} W_{(i)}^\top) \Sigma_x^{1/2} & \Sigma_y \end{bmatrix} \quad i = 1, 2,$$

Let $\mathbb{P}_{(i)}$ denote the distribution of a random i.i.d. sample of size n from $N(0, \Sigma_{(i)})$. If we further assume

$$[U_{(1)}, W_{(1)}] \begin{bmatrix} V_{(1)}^\top \\ Z_{(1)}^\top \end{bmatrix} = [U_{(2)}, W_{(2)}] \begin{bmatrix} V_{(2)}^\top \\ Z_{(2)}^\top \end{bmatrix}, \quad (5.5)$$

Then one can show that

$$D(\mathbb{P}_{(1)} || \mathbb{P}_{(2)}) = \frac{n\Delta^2(1 + \lambda_1\lambda_2)}{2(1 - \lambda_1^2)(1 - \lambda_2^2)} \|U_{(1)} V_{(1)}^\top - U_{(2)} V_{(2)}^\top\|_F^2.$$

The construction in (5.5) is crucial to prove the lower bound. The proof of this lemma can be found in Section 6.3.

5.8.2 Packing Number and Fano's Lemma

The following result on the packing number is based on the metric entropy of the Grassmannian manifold $G(k, r)$ due to Szarek [1982]. We use the version adapted from Lemma 1 of Cai et al. [2013] which is also used in Gao et al. [2015].

Lemma 5.4 For any fixed $U_0 \in \mathcal{O}(p, k)$ and $\mathcal{B}_{\epsilon_0} = \{U \in \mathcal{O}(p, k) : \|UU^\top - U_0 U_0^\top\|_F \leq \epsilon_0\}$ with $\epsilon_0 \in (0, \sqrt{2[k \wedge (p - k)]})$. Define the semi-metric $\rho(\cdot, \cdot)$ on \mathcal{B}_{ϵ_0} by

$$\rho(U_1, U_2) = \|U_1 U_1^\top - U_2 U_2^\top\|_F.$$

Then there exists universal constant C such that for any $\alpha \in (0, 1)$, the packing number $\mathcal{M}(\mathcal{B}_{\epsilon_0}, \rho, \alpha\epsilon_0)$ satisfies

$$\mathcal{M}(\mathcal{B}_{\epsilon_0}, \rho, \alpha\epsilon_0) \geq \left(\frac{1}{C\alpha}\right)^{k(p-k)}.$$

The following corollary is used to prove the lower bound.

Corollary 5.5 If we change the set in Lemma 5.4 to $\tilde{\mathcal{B}}_{\epsilon_0} = \{U \in \mathcal{O}(p, k) : \|U - U_0\|_F \leq \epsilon_0\}$, then we still have

$$\mathcal{M}(\tilde{\mathcal{B}}_{\epsilon_0}, \rho, \alpha\epsilon_0) \geq \left(\frac{1}{C\alpha}\right)^{k(p-k)}.$$

Proof Apply Lemma 5.4 to \mathcal{B}_{ϵ_0} , there exists $\mathbf{U}_1, \dots, \mathbf{U}_n$ with $n \geq (1/C\alpha)^{k(p-k)}$ such that

$$\|\mathbf{U}_i \mathbf{U}_i^\top - \mathbf{U}_0 \mathbf{U}_0^\top\|_F \leq \epsilon_0, \quad 1 \leq i \leq n, \quad \|\mathbf{U}_i \mathbf{U}_i^\top - \mathbf{U}_j \mathbf{U}_j^\top\|_F \geq \alpha \epsilon_0, \quad 1 \leq i \leq j \leq n.$$

Define $\tilde{\mathbf{U}}_i = \arg \min_{\mathbf{U} \in \{\mathbf{U}_i \mathbf{Q}, \mathbf{Q} \in \mathcal{O}(k)\}} \|\mathbf{U} - \mathbf{U}_0\|_F$, by Lemma 5.7,

$$\|\tilde{\mathbf{U}}_i - \mathbf{U}_0\|_F \leq \|\tilde{\mathbf{U}}_i \tilde{\mathbf{U}}_i^\top - \mathbf{U}_0 \mathbf{U}_0^\top\|_F \leq \epsilon_0.$$

Therefore, $\tilde{\mathbf{U}}_1, \dots, \tilde{\mathbf{U}}_n \in \tilde{\mathcal{B}}_{\epsilon_0}$ and

$$\|\tilde{\mathbf{U}}_i \tilde{\mathbf{U}}_i^\top - \tilde{\mathbf{U}}_j \tilde{\mathbf{U}}_j^\top\|_F = \|\mathbf{U}_i \mathbf{U}_i^\top - \mathbf{U}_j \mathbf{U}_j^\top\|_F \geq \alpha \epsilon_0.$$

which implies,

$$\mathcal{M}(\tilde{\mathcal{B}}_{\epsilon_0}, \rho, \alpha \epsilon_0) \geq n \geq \left(\frac{1}{C\alpha} \right)^{k(p-k)}.$$

■

Lemma 5.6 (Fano's Lemma Yu [1997]) *Let (Θ, ρ) be a (semi)metric space and $\{\mathbb{P}_\theta : \theta \in \Theta\}$ a collection of probability measures. For any totally bounded $T \subset \Theta$, denote $\mathcal{M}(T, \rho, \epsilon)$ the ϵ -packing number of T with respect to the metric ρ , i.e., the maximal number of points in T whose pairwise minimum distance in ρ is at least ϵ . Define the Kullback-Leibler diameter of T by*

$$d_{KL}(T) = \sup_{\theta, \theta' \in T} D(\mathbb{P}_\theta \| \mathbb{P}_{\theta'}).$$

Then,

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\rho^2(\hat{\theta}, \theta) \right] \geq \sup_{T \subset \Theta} \sup_{\epsilon > 0} \frac{\epsilon^2}{4} \left(1 - \frac{d_{KL}(T) + \log 2}{\log \mathcal{M}(T, \rho, \epsilon)} \right)$$

5.8.3 Proof of Lower Bound

For any fixed $[\mathbf{U}_{(0)}, \mathbf{W}_{(0)}] \in \mathcal{O}(p_1, p_1)$ and $[\mathbf{V}_{(0)}, \mathbf{Z}_{(0)}] \in \mathcal{O}(p_2, p_1)$ where $\mathbf{U}_{(0)} \in \mathbb{R}^{p_1 \times k}$, $\mathbf{V}_{(0)} \in \mathbb{R}^{p_2 \times k}$, $\mathbf{W}_{(0)} \in \mathbb{R}^{p_1 \times (p_1 - k)}$, $\mathbf{Z}_{(0)} \in \mathbb{R}^{p_2 \times (p_2 - k)}$, define

$$\mathcal{H}_{\epsilon_0} = \left\{ (\mathbf{U}, \mathbf{W}, \mathbf{V}, \mathbf{Z}) : [\mathbf{U}, \mathbf{W}] \in \mathcal{O}(p_1, p_1) \text{ with } \mathbf{U} \in \mathbb{R}^{p_1 \times k}, [\mathbf{V}, \mathbf{Z}] \in \mathcal{O}(p_2, p_1) \right.$$

$$\left. \text{with } \mathbf{V} \in \mathbb{R}^{p_2 \times k}, \|\mathbf{U} - \mathbf{U}_{(0)}\|_F \leq \epsilon_0, [\mathbf{U}, \mathbf{W}] \begin{bmatrix} \mathbf{V}^\top \\ \mathbf{Z}^\top \end{bmatrix} = [\mathbf{U}_{(0)}, \mathbf{W}_{(0)}] \begin{bmatrix} \mathbf{V}_{(0)}^\top \\ \mathbf{Z}_{(0)}^\top \end{bmatrix} \right\}.$$

For any fixed $\Sigma_x \in \mathbb{S}_+^{p_1}$, $\Sigma_y \in \mathbb{S}_+^{p_2}$ with $\kappa(\Sigma_x) = \kappa_x$, $\kappa(\Sigma_y) = \kappa_y$, consider the parametrization $\Sigma_{xy} = \Sigma_x \Phi \Lambda \Psi^\top \Sigma_y$, for $0 \leq \lambda_{k+1} < \lambda_k < 1$, define

$$\mathcal{T}_{\epsilon_0} = \left\{ \Sigma = \begin{bmatrix} \Sigma_x & \Sigma_x^{1/2}(\lambda_k \mathbf{U} \mathbf{V}^\top + \lambda_{k+1} \mathbf{W} \mathbf{Z}^\top) \Sigma_y^{1/2} \\ \Sigma_y^{1/2}(\lambda_k \mathbf{V} \mathbf{U}^\top + \lambda_{k+1} \mathbf{Z} \mathbf{W}^\top) \Sigma_x^{1/2} & \Sigma_y \end{bmatrix}, \right.$$

$$\left. \Phi = \Sigma_x^{-1/2} [\mathbf{U}, \mathbf{W}], \Psi = \Sigma_y^{-1/2} [\mathbf{V}, \mathbf{Z}], (\mathbf{U}, \mathbf{W}, \mathbf{V}, \mathbf{Z}) \in \mathcal{H}_{\epsilon_0} \right\}.$$

It is straightforward to verify that $\mathcal{T}_{\epsilon_0} \subset \mathcal{F}(p_1, p_2, k, \lambda_k, \lambda_{k+1}, \kappa_x, \kappa_y)$. For any $\Sigma_{(i)} \in \mathcal{T}_{\epsilon_0}$, $i = 1, 2$, they yield to the parametrization,

$$\Sigma_{(i)} = \begin{bmatrix} \Sigma_x & \Sigma_x^{1/2}(\lambda_k \mathbf{U}_{(i)} \mathbf{V}_{(i)}^\top + \lambda_{k+1} \mathbf{W}_{(i)} \mathbf{Z}_{(i)}^\top) \Sigma_y^{1/2} \\ \Sigma_y^{1/2}(\lambda_k \mathbf{V}_{(i)} \mathbf{U}_{(i)}^\top + \lambda_{k+1} \mathbf{Z}_{(i)} \mathbf{W}_{(i)}^\top) \Sigma_x^{1/2} & \Sigma_y \end{bmatrix},$$

where $(\mathbf{U}_{(i)}, \mathbf{W}_{(i)}, \mathbf{V}_{(i)}, \mathbf{Z}_{(i)}) \in \mathcal{H}_{\epsilon_0}$ and the leading- k canonical vectors are $\Phi_{1:k}^{(i)} = \Sigma_x^{-1/2} \mathbf{U}_{(i)}$, $\Psi_{1:k}^{(i)} = \Sigma_y^{-1/2} \mathbf{V}_{(i)}$. We define a semi-metric on \mathcal{T}_{ϵ_0} as

$$\rho(\Sigma_{(1)}, \Sigma_{(2)}) = \left\| P_{\Sigma_x^{1/2} \Phi_{1:k}^{(1)}} - P_{\Sigma_x^{1/2} \Phi_{1:k}^{(2)}} \right\|_F = \left\| P_{U_{(1)}} - P_{U_{(2)}} \right\|_F.$$

By Lemma 5.3,

$$D(\mathbb{P}_{\Sigma_1} \| \mathbb{P}_{\Sigma_2}) = \frac{n\Delta^2(1 + \lambda_k \lambda_{k+1})}{2(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)} \left\| \mathbf{U}_{(1)} \mathbf{V}_{(1)}^\top - \mathbf{U}_{(2)} \mathbf{V}_{(2)}^\top \right\|_F^2.$$

Further by the definition of $d_{KL}(T)$,

$$d_{KL}(T) = \frac{n\Delta^2(1 + \lambda_k \lambda_{k+1})}{2(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)} \sup_{\Sigma_{(1)}, \Sigma_{(2)} \in \mathcal{T}_{\epsilon_0}} \left\| \mathbf{U}_{(1)} \mathbf{V}_{(1)}^\top - \mathbf{U}_{(2)} \mathbf{V}_{(2)}^\top \right\|_F^2. \quad (5.6)$$

To bound the Kullback-Leibler diameter, for any $\Sigma_{(1)}, \Sigma_{(2)} \in \mathcal{T}_{\epsilon_0}$, by definition,

$$[\mathbf{U}_{(1)}, \mathbf{W}_{(1)}] \begin{bmatrix} \mathbf{V}_{(1)}^\top \\ \mathbf{Z}_{(1)}^\top \end{bmatrix} = [\mathbf{U}_{(2)}, \mathbf{W}_{(2)}] \begin{bmatrix} \mathbf{V}_{(2)}^\top \\ \mathbf{Z}_{(2)}^\top \end{bmatrix},$$

which implies that they are singular value decompositions of the same matrix. Therefore, there exists $\mathbf{Q} \in \mathcal{O}(p_1, p_1)$ such that

$$[\mathbf{U}_{(2)}, \mathbf{W}_{(2)}] = [\mathbf{U}_{(1)}, \mathbf{W}_{(1)}] \mathbf{Q}, \quad [\mathbf{V}_{(2)}, \mathbf{Z}_{(2)}] = [\mathbf{V}_{(1)}, \mathbf{Z}_{(1)}] \mathbf{Q}. \quad (5.7)$$

Decompose \mathbf{Q} into four blocks such that

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix}.$$

Substitute into (5.7),

$$\mathbf{U}_{(2)} = \mathbf{U}_{(1)} \mathbf{Q}_{11} + \mathbf{W}_{(1)} \mathbf{Q}_{21}, \quad \mathbf{V}_{(2)} = \mathbf{V}_{(1)} \mathbf{Q}_{11} + \mathbf{Z}_{(1)} \mathbf{Q}_{21}.$$

Then,

$$\begin{aligned} \left\| \mathbf{U}_{(2)} - \mathbf{U}_{(1)} \right\|_F^2 &= \left\| \mathbf{U}_{(1)} (\mathbf{Q}_{11} - \mathbf{I}_k) + \mathbf{W}_{(1)} \mathbf{Q}_{21} \right\|_F^2 \\ &= \left\| \mathbf{U}_{(1)} (\mathbf{Q}_{11} - \mathbf{I}_k) \right\|_F^2 + \left\| \mathbf{W}_{(1)} \mathbf{Q}_{21} \right\|_F^2 \\ &= \left\| \mathbf{Q}_{11} - \mathbf{I}_k \right\|_F^2 + \left\| \mathbf{Q}_{21} \right\|_F^2. \end{aligned}$$

The second equality is due to the fact that $\mathbf{U}_{(1)}$ and $\mathbf{W}_{(1)}$ have orthogonal column space and the third equality is valid because $\mathbf{U}_{(1)}, \mathbf{W}_{(1)} \in \mathcal{O}(p_1, k)$. By the same argument, we will have

$$\|\mathbf{V}_{(2)} - \mathbf{V}_{(1)}\|_F^2 = \|\mathbf{Q}_{11} - \mathbf{I}_k\|_F^2 + \|\mathbf{Q}_{21}\|_F^2.$$

Notice that

$$\begin{aligned} \|\mathbf{U}_{(1)} \mathbf{V}_{(1)}^\top - \mathbf{U}_{(2)} \mathbf{V}_{(2)}^\top\|_F^2 &= \|(\mathbf{U}_{(1)} - \mathbf{U}_{(2)}) \mathbf{V}_{(1)} + \mathbf{U}_{(2)} (\mathbf{V}_{(1)} - \mathbf{V}_{(2)})\|_F^2 \\ &\leq 2\|\mathbf{U}_{(1)} - \mathbf{U}_{(2)}\|_F^2 + 2\|\mathbf{V}_{(1)} - \mathbf{V}_{(2)}\|_F^2 \\ &= 4\|\mathbf{U}_{(1)} - \mathbf{U}_{(2)}\|_F^2 \\ &\leq 8(\|\mathbf{U}_{(1)} - \mathbf{U}_{(0)}\|_F^2 + \|\mathbf{U}_{(0)} - \mathbf{U}_{(2)}\|_F^2) \\ &\leq 16\epsilon_0^2. \end{aligned}$$

Then, substitute into (5.6)

$$d_{KL}(T) \leq \frac{8n\Delta^2(1 + \lambda_k\lambda_{k+1})}{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)} \epsilon_0^2. \quad (5.8)$$

Let $\mathcal{B}_{\epsilon_0} = \{\mathbf{U} \in \mathcal{O}(p_1, k) : \|\mathbf{U} - \mathbf{U}_{(0)}\|_F \leq \epsilon_0\}$. Under the semi-metric $\tilde{\rho}(\mathbf{U}_{(1)}, \mathbf{U}_{(2)}) = \|\mathbf{U}_{(1)} \mathbf{U}_{(1)}^\top - \mathbf{U}_{(2)} \mathbf{U}_{(2)}^\top\|_F$, we claim that the packing number of \mathcal{H}_{ϵ_0} is lower bounded by the packing number of \mathcal{B}_{ϵ_0} . To prove this claim, it suffices to show that for any $\mathbf{U} \in \mathcal{B}_{\epsilon_0}$, there exists corresponding $\mathbf{W}, \mathbf{V}, \mathbf{Z}$ such that $(\mathbf{U}, \mathbf{W}, \mathbf{V}, \mathbf{Z}) \in \mathcal{H}_{\epsilon_0}$. First of all, by definition, $\|\mathbf{U} - \mathbf{U}_0\|_F \leq \epsilon_0$. Let $\mathbf{W} \in \mathcal{O}(p_1, p_1 - k)$ be the orthogonal complement of \mathbf{U} . Then $[\mathbf{U}, \mathbf{W}] \in \mathcal{O}(p_1, p_1)$ and therefore there exists $\mathbf{Q} \in \mathcal{O}(p_1, p_1)$ such that

$$[\mathbf{U}, \mathbf{W}] = [\mathbf{U}_{(0)}, \mathbf{W}_0] \mathbf{Q}.$$

Set $[\mathbf{V}, \mathbf{Z}] = [\mathbf{V}_{(0)}, \mathbf{Z}_0] \mathbf{Q} \in \mathcal{O}(p_2, p_1)$, then

$$[\mathbf{U}, \mathbf{W}] \begin{bmatrix} \mathbf{V}^\top \\ \mathbf{Z}^\top \end{bmatrix} = [\mathbf{U}_{(0)}, \mathbf{W}_{(0)}] \begin{bmatrix} \mathbf{V}_{(0)}^\top \\ \mathbf{Z}_{(0)}^\top \end{bmatrix},$$

which implies $(\mathbf{U}, \mathbf{W}, \mathbf{V}, \mathbf{Z}) \in \mathcal{H}_{\epsilon_0}$. Let

$$\epsilon = \alpha \epsilon_0 = c \left(\sqrt{k \wedge (p_1 - k)} \wedge \sqrt{\frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{n\Delta^2(1 + \lambda_k\lambda_{k+1})} k(p_1 - k)} \right),$$

where $c \in (0, 1)$ depends on α and is chosen small enough such that $\epsilon_0 = \epsilon/\alpha \in (0, \sqrt{2[k \wedge (p_1 - k)]})$. By Corollary 5.5,

$$\mathcal{M}(\mathcal{T}_{\epsilon_0}, \rho, \alpha \epsilon_0) = \mathcal{M}(\mathcal{H}_{\epsilon_0}, \tilde{\rho}, \alpha \epsilon_0) \geq \mathcal{M}(\mathcal{B}_{\epsilon_0}, \tilde{\rho}, \alpha \epsilon_0) \geq \left(\frac{1}{C\alpha} \right)^{k(p_1 - k)}.$$

Apply Lemma 5.6 with $\mathcal{T}_{\epsilon_0}, \rho, \epsilon$,

$$\inf_{\hat{\Phi}_{1:k}} \sup_{\Sigma \in \mathcal{F}} \mathbb{E} \left[\left\| \mathbf{P}_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - \mathbf{P}_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|_F^2 \right] \geq \sup_{T \subset \Theta} \sup_{\epsilon > 0} \frac{\epsilon^2}{4} \left(1 - \frac{8c^2 k(p_1 - k) + \log 2}{k(p_1 - k) \log \frac{1}{C\alpha}} \right).$$

Choose α small enough such that

$$1 - \frac{8c^2 k(p_1 - k) + \log 2}{k(p_1 - k) \log \frac{1}{C\alpha}} \geq \frac{1}{2}.$$

Then the lower bound is reduced to

$$\begin{aligned} \inf_{\hat{\Phi}_{1:k}} \sup_{\Sigma \in \mathcal{F}} \mathbb{E} \left[\left\| P_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|_F^2 \right] &\geq \frac{c^2}{8} \left\{ \frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{n\Delta^2(1 + \lambda_k \lambda_{k+1})} k(p_1 - k) \wedge k \wedge (p_1 - k) \right\} \\ &\geq C^2 k \left\{ \left(\frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{\Delta^2} \frac{p_1 - k}{n} \right) \wedge 1 \wedge \frac{p_1 - k}{k} \right\} \end{aligned}$$

By symmetry,

$$\inf_{\hat{\Psi}_{1:k}} \sup_{\Sigma \in \mathcal{F}} \mathbb{E} \left[\left\| P_{\Sigma_y^{1/2} \hat{\Psi}_{1:k}} - P_{\Sigma_y^{1/2} \Psi_{1:k}} \right\|_F^2 \right] \geq C^2 k \left\{ \left(\frac{(1 - \lambda_k^2)(1 - \lambda_{k+1}^2)}{\Delta^2} \frac{p_1 - k}{n} \right) \wedge 1 \wedge \frac{p_1 - k}{k} \right\}$$

The lower bound for operator norm error can be immediately obtained by noticing that $P_{\Sigma_y^{1/2} \hat{\Psi}_{1:k}} - P_{\Sigma_y^{1/2} \Psi_{1:k}}$ has at most rank $2k$ and

$$\left\| P_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|^2 \geq \frac{1}{2k} \left\| P_{\Sigma_x^{1/2} \hat{\Phi}_{1:k}} - P_{\Sigma_x^{1/2} \Phi_{1:k}} \right\|_F^2$$

5.9 Technique Lemmas

Lemma 5.7 For any matrices $\mathbf{U}_1, \mathbf{U}_2 \in \mathcal{O}(p, k)$,

$$\inf_{\mathbf{Q} \in \mathcal{O}(k, k)} \left\| \mathbf{U}_1 - \mathbf{U}_2 \mathbf{Q} \right\|_F \leq \left\| P_{\mathbf{U}_1} - P_{\mathbf{U}_2} \right\|_F$$

See Section 6.4 for the proof.

Lemma 5.8 (Wedin [1983]) For $\mathbf{U}_i \in \mathcal{O}(p, k)$ and $P_i = \mathbf{U}_i \mathbf{U}_i^\top$, $i = 1, 2$,

$$\begin{aligned} \|P_1 - P_2\|_F^2 &= 2\|(I_p - P_1)P_2\|_F^2 = 2\|(I_p - P_2)P_1\|_F^2, \\ \|P_1 - P_2\|^2 &= \|(I_p - P_1)P_2\|^2 = \|(I_p - P_2)P_1\|^2 = 1 - \sigma_{\min}^2(P_1 P_2) \end{aligned}$$

The following variant of Wedin's $\sin \theta$ law [Wedin, 1972] is proved in Proposition 1 of Cai et al. [2015]

Lemma 5.9 For $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{m \times n}$ and $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{E}$, define the singular value decomposition of \mathbf{A} and $\hat{\mathbf{A}}$ as

$$\mathbf{A} = \mathbf{U} \mathbf{D} \mathbf{V}^\top, \quad \hat{\mathbf{A}} = \hat{\mathbf{U}} \hat{\mathbf{D}} \hat{\mathbf{V}}^\top.$$

Then the following perturbation bound holds,

$$\left\| (\mathbf{I} - P_{\mathbf{U}_{1:k}}) P_{\hat{\mathbf{U}}_{1:k}} \right\| = \left\| P_{\mathbf{U}_{1:k}} - P_{\hat{\mathbf{U}}_{1:k}} \right\| \leq \frac{2\|\mathbf{E}\|}{\sigma_k(\mathbf{A}) - \sigma_{k+1}(\mathbf{A})},$$

where $\sigma_k(\mathbf{A}), \sigma_{k+1}(\mathbf{A})$ are the k th and $(k+1)$ th singular values of \mathbf{A} respectively.

Lemma 5.10 (Covariance Matrix Estimation, [Vershynin \[2010\]](#), Remark 5.40) Assume $\mathbf{A} \in \mathbb{R}^{n \times p}$ has independent sub-gaussian random rows with second moment matrix Σ . Then there exists universal constant C such that for every $t \geq 0$, the following inequality holds with probability at least $1 - e^{-ct^2}$,

$$\left\| \frac{1}{n} \mathbf{A}^\top \mathbf{A} - \Sigma \right\| \leq \max\{\delta, \delta^2\} \|\Sigma\| \quad \delta = C \sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}. \quad (5.9)$$

Lemma 5.11 Assume $\mathbf{A} \in \mathbb{R}^{n \times p}$, $n \geq p$ has independent sub-gaussian random rows with second moment matrix Σ . Then there exists universal constant C such that

$$\mathbb{E} \left\| \frac{1}{n} \mathbf{A}^\top \mathbf{A} - \Sigma \right\|^4 \leq C \frac{p^2}{n^2} \|\Sigma\|^4. \quad (5.10)$$

See Section 6.5 for the proof.

Lemma 5.12 (Bernstein inequality, [Vershynin \[2010\]](#), Proposition 5.16) Let X_1, \dots, X_n be independent centered sub-exponential random variables and $K = \max_i \|X_i\|_{\psi_1}$. Then for every $\mathbf{a} = (\mathbf{a}_1, \dots, \mathbf{a}_n) \in \mathbb{R}^n$ and every $t \geq 0$, we have

$$P \left\{ \left| \sum_{i=1}^n \mathbf{a}_i X_i \right| \geq t \right\} \leq 2 \exp \left\{ -c \min \left(\frac{t^2}{K^2 \|\mathbf{a}\|_2^2}, \frac{t}{K \|\mathbf{a}\|_\infty} \right) \right\}. \quad (5.11)$$

Lemma 5.13 (Hanson-Wright inequality, [Rudelson and Vershynin \[2013\]](#)) Let $\mathbf{x} = (x_1, \dots, x_p)$ be a random vectors with independent components x_i which satisfy $\mathbb{E} x_i = 0$ and $\|x_i\|_{\psi_2} \leq K$, Let $\mathbf{A} \in \mathbb{R}^{p \times p}$. Then there exists universal constant c such that for every $t \geq 0$,

$$P \left\{ |\mathbf{x}^\top \mathbf{A} \mathbf{x} - \mathbb{E} \mathbf{x}^\top \mathbf{A} \mathbf{x}| \geq t \right\} \leq 2 \exp \left\{ -c \min \left(\frac{t^2}{K^4 \|\mathbf{A}\|_F^2}, \frac{t}{K^2 \|\mathbf{A}\|_2} \right) \right\}. \quad (5.12)$$

Lemma 5.14 (Covering Number of the Sphere, [Vershynin \[2010\]](#), Lemma 5.2). The unit Euclidean sphere \mathbb{S}^{n-1} equipped with the Euclidean metric satisfies for every $\epsilon > 0$ that

$$|\mathcal{N}(\mathbb{S}^{n-1}, \epsilon)| \leq \left(1 + \frac{2}{\epsilon}\right)^n, \quad (5.13)$$

where $\mathcal{N}(\mathbb{S}^{n-1}, \epsilon)$ is the ϵ -net of \mathbb{S}^{n-1} with minimal cardinality.

6 Appendix

6.1 Proof of Lemma 4.5

In this section, we show how to control the higher order terms δ_6 and δ_7 . The universal constants C, C_1, c, \dots might change from line to line. To facilitate presentation, we again introduce $\mathbf{z} = (\mathbf{x}^\top, \mathbf{y}^\top)^\top$ as the concatenation of \mathbf{x} and \mathbf{y} . Then

$$\Sigma_z = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} = \begin{bmatrix} I_{p_1} & \Sigma_{xy} \\ \Sigma_{yx} & I_{p_2} \end{bmatrix}, \quad \hat{\Sigma}_z = \begin{bmatrix} \hat{\Sigma}_x & \hat{\Sigma}_{xy} \\ \hat{\Sigma}_{yx} & \hat{\Sigma}_y \end{bmatrix}.$$

Lemma 6.1 *There exists universal constant C such that the following inequalities hold deterministically*

$$\begin{aligned} & \|\hat{\Sigma}_x\|, \|\hat{\Sigma}_y\|, \|\hat{\Sigma}_{xy}\| \leq \|\hat{\Sigma}_z\|, \\ & \|\Sigma_x - \hat{\Sigma}_x\|, \|\Sigma_y - \hat{\Sigma}_y\|, \|\Sigma_{xy} - \hat{\Sigma}_{xy}\| \leq \|\Sigma_z - \hat{\Sigma}_z\|, \\ & \|\hat{\Lambda} - \Lambda\| \leq \|\hat{\Sigma}_x^{-1/2} \hat{\Sigma}_{xy} \hat{\Sigma}_y^{-1/2} - \Sigma_{xy}\| \leq \|\Sigma_z - \hat{\Sigma}_z\| (2 + \|\hat{\Sigma}_z\|), \\ & \|\hat{\Phi}_{1:k}^l\|, \|\hat{\Psi}_{1:k}^l\| \leq C \|\Sigma_z - \hat{\Sigma}_z\| \left(\frac{2 + \|\hat{\Sigma}_z\|}{\Delta} + \|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\| \right), \end{aligned}$$

where $\Delta = \lambda_k - \lambda_{k+1}$ is the eigen-gap.

Lemma 6.1 and triangle inequality are frequently employed in this proof. Notice that $\Sigma_x^{ij}, \Sigma_y^{ij}, \Sigma_{xy}^{ij}, \Sigma_{yx}^{ij}, 1 \leq i, j \leq 2$ are sub-matrices of $\Sigma_x, \Sigma_y, \Sigma_{xy}, \Sigma_{yx}$. We will also repeatedly use the fact that the operator norm of a matrix is no less than that of its sub-matrices.

$$\begin{aligned} \|\delta_1\| & \leq \|\hat{\Phi}_{1:k}^u (\hat{\Lambda}_1 - \Lambda_1) + (\hat{\Sigma}_x^{11} - I_k) \hat{\Phi}_{1:k}^u \hat{\Lambda}_1 - (\hat{\Sigma}_{xy}^{11} - \Lambda_1) \hat{\Psi}_{1:k}^u + \hat{\Sigma}_x^{12} \hat{\Phi}_{1:k}^l \hat{\Lambda}_1 - \hat{\Sigma}_{xy}^{12} \hat{\Psi}_{1:k}^l\| \\ & \leq \|\hat{\Phi}_{1:k}\| \|\Sigma_z - \hat{\Sigma}_z\| (2 + \|\hat{\Sigma}_z\|) + \|\Sigma_z - \hat{\Sigma}_z\| \|\hat{\Phi}_{1:k}\| + \|\Sigma_z - \hat{\Sigma}_z\| \|\hat{\Psi}_{1:k}\| \\ & \quad + 2C \|\Sigma_z - \hat{\Sigma}_z\|^2 \left(\frac{2 + \|\hat{\Sigma}_z\|}{\Delta} + \|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\| \right) \\ & \leq C_1 \|\Sigma_z - \hat{\Sigma}_z\| (2 + \|\hat{\Sigma}_z\|) (\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|) + C_2 \|\Sigma_z - \hat{\Sigma}_z\|^2 \left(\frac{2 + \|\hat{\Sigma}_z\|}{\Delta} + \|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\| \right) \\ & \leq C_3 \|\Sigma_z - \hat{\Sigma}_z\| (2 + \|\hat{\Sigma}_z\|) (\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|) (1 + \|\Sigma_z - \hat{\Sigma}_z\|/\Delta). \end{aligned}$$

where in the last inequality we use $\|\Sigma_z\| \leq 2$ and

$$\|\Sigma_z - \hat{\Sigma}_z\|^2 (\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|) \leq \|\Sigma_z - \hat{\Sigma}_z\| (\|\Sigma_z\| + \|\hat{\Sigma}_z\|) (\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|)$$

By the same argument,

$$\|\delta_2\| \leq C_3 \|\Sigma_z - \hat{\Sigma}_z\| (2 + \|\hat{\Sigma}_z\|) (\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|) (1 + \|\Sigma_z - \hat{\Sigma}_z\|/\Delta).$$

We can also bound δ_3, δ_4 in the same manner and will obtain,

$$\|\delta_3\|, \|\delta_4\| \leq C\|\Sigma_z - \hat{\Sigma}_z\|^2(2 + \|\hat{\Sigma}_z\|)^2(\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|)/\Delta$$

Recall that $\delta_6 = \hat{\Sigma}_x^{21}\delta_1\Lambda_1 - \Lambda_2\hat{\Sigma}_y^{21}\delta_1 + \delta_5$ and $\delta_5 = -\delta_3\Lambda_1 + \Lambda_2\delta_4$, then

$$\begin{aligned} \|\mathbf{A}_1 \circ \delta_6\| &\leq \|\mathbf{A}_1 \circ (\hat{\Sigma}_x^{21}\delta_1\Lambda_1 - \Lambda_2\hat{\Sigma}_y^{21}\delta_1 + \delta_5)\| \\ &\leq \|\mathbf{A}_1 \circ (\hat{\Sigma}_x^{21}\delta_1\Lambda_1)\| + \|\mathbf{A}_1 \circ (\Lambda_2\hat{\Sigma}_y^{21}\delta_1)\| + \|\mathbf{A}_1 \circ (\delta_3\Lambda_1)\| + \|\mathbf{A}_1 \circ (\Lambda_2\delta_4)\| \\ &= \|(\mathbf{A}_1\Lambda_1) \circ (\hat{\Sigma}_x^{21}\delta_1)\| + \|(\Lambda_2\mathbf{A}_1) \circ (\hat{\Sigma}_y^{21}\delta_1)\| + \|(\mathbf{A}_1\Lambda_1) \circ \delta_3\| + \|(\Lambda_2\mathbf{A}_1) \circ \delta_4\|. \end{aligned}$$

By the same argument as in (4.16),

$$\begin{aligned} \|\mathbf{A}_1 \circ \delta_6\| &\leq \frac{3}{2}\|\hat{\Sigma}_x^{21}\delta_1\| + \frac{1}{2}\|\hat{\Sigma}_y^{21}\delta_1\| + \frac{3}{2}\|\delta_3\| + \frac{1}{2}\|\delta_4\| \\ &\leq 2\|\Sigma_z - \hat{\Sigma}_z\|\|\delta_1\| + \frac{3}{2}\|\delta_3\| + \frac{1}{2}\|\delta_4\| \\ &\leq C\|\Sigma_z - \hat{\Sigma}_z\|^2(2 + \|\hat{\Sigma}_z\|)(\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \hat{\Sigma}_z\|/\Delta) \\ &\quad + C\|\Sigma_z - \hat{\Sigma}_z\|^2(2 + \|\hat{\Sigma}_z\|)^2(\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|)/\Delta \\ &\leq C\|\Sigma_z - \hat{\Sigma}_z\|^2(2 + \|\hat{\Sigma}_z\|)^2(\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \hat{\Sigma}_z\|/\Delta). \end{aligned}$$

By the same argument as in (4.17),

$$\begin{aligned} \|\mathbf{A}_1 \circ \mathbf{A}_2 \circ \delta_6\| &\leq \frac{1}{\Delta}\|\mathbf{A}_1 \circ \delta_6\| \\ &\leq C\|\Sigma_z - \hat{\Sigma}_z\|^2(2 + \|\hat{\Sigma}_z\|)^2(\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \hat{\Sigma}_z\|/\Delta)^2. \end{aligned}$$

Recall that $\delta_7 = \hat{\Sigma}_{xy}^{21}\delta_1 + \hat{\Sigma}_x^{21}(\Lambda_1\delta_1 + \delta_2\Lambda_1) + \Lambda_2\hat{\Sigma}_y^{21}\delta_2 + \delta_5$, then

$$\begin{aligned} \|\delta_7\| &\leq \|\Sigma_z - \hat{\Sigma}_z\|(2\|\delta_1\| + 2\|\delta_2\|) + \|\delta_3\| + \|\delta_4\| \\ &\leq C\|\Sigma_z - \hat{\Sigma}_z\|^2(2 + \|\hat{\Sigma}_z\|)(\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \hat{\Sigma}_z\|/\Delta) \\ &\quad + C\|\Sigma_z - \hat{\Sigma}_z\|^2(2 + \|\hat{\Sigma}_z\|)^2(\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|)/\Delta \\ &\leq C\|\Sigma_z - \hat{\Sigma}_z\|^2(2 + \|\hat{\Sigma}_z\|)^2(\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \hat{\Sigma}_z\|/\Delta). \end{aligned}$$

Again, by the same argument in (4.17),

$$\|\mathbf{A}_2 \circ \delta_7\| \leq C\|\Sigma_z - \hat{\Sigma}_z\|^2(2 + \|\hat{\Sigma}_z\|)^2(\|\hat{\Phi}_{1:k}\| + \|\hat{\Psi}_{1:k}\|)(1 + \|\Sigma_z - \hat{\Sigma}_z\|/\Delta)^2.$$

6.2 Proof of Lemma 6.1

The first two inequalities are trivial because the operator norm of a matrix is not less than that of its sub-matrices. Notice that $\hat{\Lambda}$ and Λ are singular values of $\hat{\Sigma}_x^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_y^{-1/2}$ and Σ_{xy} respectively. Hence by Weyl's inequality,

$$\|\hat{\Lambda} - \Lambda\| \leq \|\hat{\Sigma}_x^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_y^{-1/2} - \Sigma_{xy}\|.$$

Further observe that

$$\begin{aligned}\hat{\Sigma}_x^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_y^{-1/2} - \Sigma_{xy} &= (\mathbf{I}_{p_1} - \hat{\Sigma}_x^{1/2})\hat{\Sigma}_x^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_y^{-1/2} \\ &\quad + \hat{\Sigma}_x^{1/2}\hat{\Sigma}_x^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_y^{-1/2}(\mathbf{I}_{p_2} - \hat{\Sigma}_y^{1/2}) + (\hat{\Sigma}_{xy} - \Sigma_{xy}).\end{aligned}$$

and $\|\hat{\Sigma}_x^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_y^{-1/2}\| = \hat{\lambda}_1 \leq 1$. Then

$$\|\hat{\Sigma}_x^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_y^{-1/2} - \Sigma_{xy}\| \leq \|\mathbf{I}_{p_1} - \hat{\Sigma}_x^{1/2}\| + \|\hat{\Sigma}_x\|\|\mathbf{I}_{p_2} - \hat{\Sigma}_y^{1/2}\| + \|\hat{\Sigma}_{xy} - \Sigma_{xy}\|.$$

Also notice that

$$\|\mathbf{I} - \hat{\Sigma}_y\| = \|(\mathbf{I} - \Sigma_y^{1/2})(\mathbf{I} + \Sigma_y^{1/2})\| \geq \sigma_{\min}(\mathbf{I} + \Sigma_y^{1/2})\|\mathbf{I} - \Sigma_y^{1/2}\| \geq \|\mathbf{I} - \Sigma_y^{1/2}\|.$$

Therefore,

$$\begin{aligned}\|\hat{\Sigma}_x^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_y^{-1/2} - \Sigma_{xy}\| &\leq \|\mathbf{I}_{p_1} - \hat{\Sigma}_x\| + \|\hat{\Sigma}_x\|\|\mathbf{I}_{p_2} - \hat{\Sigma}_y\| + \|\hat{\Sigma}_{xy} - \Sigma_{xy}\| \\ &\leq \|\Sigma_z - \hat{\Sigma}_z\| \left(2 + \|\hat{\Sigma}_z\|\right).\end{aligned}$$

The last inequality in the lemma relies on the fact that $\hat{\Sigma}_x^{1/2}\hat{\Phi}_{1:k}$ and $\mathbf{I}_{p_1,k}$ are leading k singular vectors of $\hat{\Sigma}_x^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_y^{-1/2}$ and Σ_{xy} respectively. By a variant of Wedin's $\sin \theta$ law as stated in Lemma 5.9,

$$\left\|P_{\hat{\Sigma}_x^{1/2}\hat{\Phi}_{1:k}}(\mathbf{I}_{p_1} - P_{\mathbf{I}_{p_1,k}})\right\| \leq \frac{C\|\hat{\Sigma}_x^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_y^{-1/2} - \Sigma_{xy}\|}{\Delta}.$$

On the other hand,

$$\begin{aligned}\left\|P_{\hat{\Sigma}_x^{1/2}\hat{\Phi}_{1:k}}(\mathbf{I}_{p_1} - P_{\mathbf{I}_{p_1,k}})\right\| &= \left\|\hat{\Sigma}_x^{1/2}\hat{\Phi}_{1:k}(\hat{\Sigma}_x^{1/2}\hat{\Phi}_{1:k})^\top(\mathbf{I}_{p_1} - P_{\mathbf{I}_{p_1,k}})\right\| \\ &= \left\|(\hat{\Sigma}_x^{1/2}\hat{\Phi}_{1:k})^\top(\mathbf{I}_{p_1} - P_{\mathbf{I}_{p_1,k}})\right\| \\ &= \left\|(\hat{\Sigma}_x^{1/2}\hat{\Phi}_{1:k})^l\right\|,\end{aligned}$$

where the second equality is due to the fact that $\hat{\Sigma}_x^{1/2}\hat{\Phi}_{1:k}$ has orthonormal columns and $(\hat{\Sigma}_x^{1/2}\hat{\Phi}_{1:k})^l$ denotes the lower $(p_1 - k) \times k$ sub-matrix of $\hat{\Sigma}_x^{1/2}\hat{\Phi}_{1:k}$. Again, by triangle inequality,

$$\begin{aligned}\left\|\hat{\Phi}_{1:k}^l\right\| &= \left\|(\hat{\Sigma}_x^{1/2}\hat{\Phi}_{1:k})^l - \left((\hat{\Sigma}_x^{1/2} - \mathbf{I}_{p_1})\hat{\Phi}_{1:k}\right)^l\right\| \\ &\leq \left\|(\hat{\Sigma}_x^{1/2}\hat{\Phi}_{1:k})^l\right\| + \left\|(\hat{\Sigma}_x^{1/2} - \mathbf{I}_{p_1})\hat{\Phi}_{1:k}\right\| \\ &\leq \frac{C\|\hat{\Sigma}_x^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_y^{-1/2} - \Sigma_{xy}\|}{\Delta} + \|\hat{\Sigma}_z - \hat{\Sigma}\|\|\hat{\Phi}_{1:k}\|.\end{aligned}$$

The last inequality is obtained by substituting the upper bound for $\|\hat{\Sigma}_x^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_y^{-1/2} - \Sigma_{xy}\|$ obtained above.

6.3 Proof of Lemma 5.3

By simple algebra, the Kullback-Leibler divergence between two multivariate gaussian distributions satisfies

$$D(\mathbb{P}_{\Sigma_{(1)}} || \mathbb{P}_{\Sigma_{(2)}}) = \frac{n}{2} \left\{ \text{Tr} \left(\Sigma_{(2)}^{-1} (\Sigma_{(1)} - \Sigma_{(2)}) \right) - \log \det(\Sigma_{(2)}^{-1} \Sigma_{(1)}) \right\}.$$

Notice that

$$\Sigma_{(i)} = \begin{bmatrix} \Sigma_x^{1/2} & \\ & \Sigma_y^{1/2} \end{bmatrix} \Omega_{(i)} \begin{bmatrix} \Sigma_x^{1/2} & \\ & \Sigma_y^{1/2} \end{bmatrix},$$

where

$$\Omega_{(i)} = \begin{bmatrix} I_{p_1} & \lambda_1 U_{(i)} V_{(i)}^\top + \lambda_2 W_{(i)} Z_{(i)}^\top \\ \lambda_1 V_{(i)} U_{(i)}^\top + \lambda_2 Z_{(i)} W_{(i)}^\top & I_{p_2} \end{bmatrix}.$$

Then,

$$D(\mathbb{P}_{\Sigma_{(1)}} || \mathbb{P}_{\Sigma_{(2)}}) = \frac{n}{2} \left\{ \text{Tr}(\Omega_{(2)}^{-1} \Omega_{(1)}) - (p_1 + p_2) - \log \det(\Omega_{(2)}^{-1} \Omega_{(1)}) \right\}.$$

Also notice that

$$\begin{aligned} \Omega_{(i)} &= \begin{bmatrix} I_{p_1} & \\ & I_{p_2} \end{bmatrix} + \frac{\lambda_1}{2} \begin{bmatrix} U_{(i)} \\ V_{(i)} \end{bmatrix} \begin{bmatrix} U_{(i)}^\top & V_{(i)}^\top \end{bmatrix} - \frac{\lambda_1}{2} \begin{bmatrix} U_{(i)} \\ -V_{(i)} \end{bmatrix} \begin{bmatrix} U_{(i)}^\top & -V_{(i)}^\top \end{bmatrix} \\ &\quad + \frac{\lambda_2}{2} \begin{bmatrix} W_{(i)} \\ Z_{(i)} \end{bmatrix} \begin{bmatrix} W_{(i)}^\top & Z_{(i)}^\top \end{bmatrix} - \frac{\lambda_2}{2} \begin{bmatrix} W_{(i)} \\ -Z_{(i)} \end{bmatrix} \begin{bmatrix} W_{(i)}^\top & -Z_{(i)}^\top \end{bmatrix}. \end{aligned}$$

Therefore $\Omega_{(1)}, \Omega_{(2)}$ share the same set of eigenvalues: $1 + \lambda_1$ with multiplicity k , $1 - \lambda_1$ with multiplicity k , $1 + \lambda_2$ with multiplicity $p_1 - k$, $1 - \lambda_2$ with multiplicity $p_1 - k$ and 1 with multiplicity $2(p_2 - p_1)$. This implies $\log \det(\Omega_{(2)}^{-1} \Omega_{(1)}) = 0$. On the other hand, by block inversion formula, we can compute

$$\Omega_{(2)}^{-1} = \begin{bmatrix} I_{p_1} + \frac{\lambda_1^2}{1-\lambda_1^2} U_{(2)} U_{(2)}^\top + \frac{\lambda_2^2}{1-\lambda_2^2} W_{(2)} W_{(2)}^\top & -\frac{\lambda_1}{1-\lambda_1^2} U_{(2)} V_{(2)}^\top - \frac{\lambda_2}{1-\lambda_2^2} W_{(2)} Z_{(2)}^\top \\ -\frac{\lambda_1}{1-\lambda_1^2} V_{(2)} U_{(2)}^\top - \frac{\lambda_2}{1-\lambda_2^2} Z_{(2)} W_{(2)}^\top & I_{p_2} + \frac{\lambda_1^2}{1-\lambda_1^2} V_{(2)} V_{(2)}^\top + \frac{\lambda_2^2}{1-\lambda_2^2} Z_{(2)} Z_{(2)}^\top \end{bmatrix}.$$

Divide $\Omega_{(2)}^{-1} \Omega_{(1)}$ into blocks such that

$$\Omega_{(2)}^{-1} \Omega_{(1)} = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix} \quad \text{where } J_{11} \in \mathbb{R}^{p_1 \times p_1}, J_{22} \in \mathbb{R}^{p_2 \times p_2},$$

and

$$\begin{aligned} J_{11} &= \frac{\lambda_1^2}{1-\lambda_1^2} (U_{(2)} U_{(2)}^\top - U_{(2)} V_{(2)}^\top V_{(1)} U_{(1)}^\top) + \frac{\lambda_2^2}{1-\lambda_2^2} (W_{(2)} W_{(2)}^\top - W_{(2)} Z_{(2)}^\top Z_{(1)} W_{(1)}^\top) \\ &\quad - \frac{\lambda_1 \lambda_2}{1-\lambda_1^2} (U_{(2)} V_{(2)}^\top Z_{(1)} W_{(1)}^\top) - \frac{\lambda_1 \lambda_2}{1-\lambda_2^2} (W_{(2)} Z_{(2)}^\top V_{(1)} U_{(1)}^\top) \\ J_{22} &= \frac{\lambda_1^2}{1-\lambda_1^2} (V_{(2)} V_{(2)}^\top - V_{(2)} U_{(2)}^\top U_{(1)} V_{(1)}^\top) + \frac{\lambda_2^2}{1-\lambda_2^2} (Z_{(2)} Z_{(2)}^\top - Z_{(2)} W_{(2)}^\top W_{(1)} Z_{(1)}^\top) \\ &\quad - \frac{\lambda_1 \lambda_2}{1-\lambda_1^2} (V_{(2)} U_{(2)}^\top W_{(1)} Z_{(1)}^\top) - \frac{\lambda_1 \lambda_2}{1-\lambda_2^2} (Z_{(2)} W_{(2)}^\top U_{(1)} V_{(1)}^\top). \end{aligned}$$

We spell out the algebra for $\text{tr}(\mathbf{J}_{11})$, and $\text{tr}(\mathbf{J}_{22})$ can be computed in exactly the same fashion.

$$\begin{aligned}\text{tr}(\mathbf{U}_{(2)}\mathbf{U}_{(2)}^\top - \mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top\mathbf{V}_{(1)}\mathbf{U}_{(1)}^\top) &= \frac{1}{2}\text{tr}(\mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top\mathbf{V}_{(2)}\mathbf{U}_{(2)}^\top + \mathbf{U}_{(1)}\mathbf{V}_{(1)}^\top\mathbf{V}_{(1)}\mathbf{U}_{(1)}^\top - 2\mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top\mathbf{V}_{(1)}\mathbf{U}_{(1)}^\top) \\ &= \frac{1}{2}\|\mathbf{U}_{(1)}\mathbf{V}_{(1)}^\top - \mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top\|_F^2.\end{aligned}$$

Similarly,

$$\text{tr}(\mathbf{W}_{(2)}\mathbf{W}_{(2)} - \mathbf{W}_{(2)}\mathbf{Z}_{(2)}^\top\mathbf{Z}_{(1)}\mathbf{W}_{(1)}^\top) = \frac{1}{2}\|\mathbf{W}_{(1)}\mathbf{Z}_{(1)}^\top - \mathbf{W}_{(2)}\mathbf{Z}_{(2)}^\top\|_F^2.$$

By the assumption, $\mathbf{U}_{(1)}\mathbf{V}_{(1)}^\top + \mathbf{W}_{(1)}\mathbf{Z}_{(1)}^\top = \mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top + \mathbf{W}_{(2)}\mathbf{Z}_{(2)}^\top$, which implies

$$\text{tr}(\mathbf{W}_{(2)}\mathbf{W}_{(2)} - \mathbf{W}_{(2)}\mathbf{Z}_{(2)}^\top\mathbf{Z}_{(1)}\mathbf{W}_{(1)}^\top) = \frac{1}{2}\|\mathbf{U}_{(1)}\mathbf{V}_{(1)}^\top - \mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top\|_F^2.$$

Further,

$$\begin{aligned}\text{tr}(\mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top\mathbf{Z}_{(1)}\mathbf{W}_{(1)}^\top) &= \text{tr}\left(\mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top(\mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top + \mathbf{W}_{(2)}\mathbf{Z}_{(2)}^\top - \mathbf{U}_{(1)}\mathbf{V}_{(1)}^\top)^\top\right) \\ &= \text{tr}\left(\mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top(\mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top - \mathbf{U}_{(1)}\mathbf{V}_{(1)}^\top)^\top\right) \\ &= \frac{1}{2}\|\mathbf{U}_{(1)}\mathbf{V}_{(1)}^\top - \mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top\|_F^2,\end{aligned}$$

and by the same argument,

$$\text{tr}(\mathbf{W}_{(2)}\mathbf{Z}_{(2)}^\top\mathbf{V}_{(1)}\mathbf{U}_{(1)}^\top) = \frac{1}{2}\|\mathbf{U}_{(1)}\mathbf{V}_{(1)}^\top - \mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top\|_F^2.$$

Sum these equations,

$$\begin{aligned}\text{tr}(\mathbf{J}_{11}) &= \frac{1}{2}\left\{\frac{\lambda_1^2}{1-\lambda_1^2} + \frac{\lambda_2^2}{1-\lambda_2^2} - \frac{\lambda_1\lambda_2}{1-\lambda_1^2} - \frac{\lambda_1\lambda_2}{1-\lambda_2^2}\right\}\|\mathbf{U}_{(1)}\mathbf{V}_{(1)}^\top - \mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top\|_F^2 \\ &= \frac{\Delta^2(1+\lambda_1\lambda_2)}{2(1-\lambda_1^2)(1-\lambda_2^2)}\|\mathbf{U}_{(1)}\mathbf{V}_{(1)}^\top - \mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top\|_F^2.\end{aligned}$$

Repeat the argument for \mathbf{J}_{22} , one can show that

$$\text{tr}(\mathbf{J}_{22}) = \text{tr}(\mathbf{J}_{11}) = \frac{\Delta^2(1+\lambda_1\lambda_2)}{2(1-\lambda_1^2)(1-\lambda_2^2)}\|\mathbf{U}_{(1)}\mathbf{V}_{(1)}^\top - \mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top\|_F^2.$$

Therefore,

$$\begin{aligned}D(\mathbb{P}_{\Sigma_{(1)}}||\mathbb{P}_{\Sigma_{(2)}}) &= \frac{n}{2}\text{tr}(\mathbf{\Omega}_{(2)}^{-1}\mathbf{\Omega}_{(1)}) = \frac{n}{2}(\text{tr}(\mathbf{J}_{11}) + \text{tr}(\mathbf{J}_{22})) \\ &= \frac{n\Delta^2(1+\lambda_1\lambda_2)}{2(1-\lambda_1^2)(1-\lambda_2^2)}\|\mathbf{U}_{(1)}\mathbf{V}_{(1)}^\top - \mathbf{U}_{(2)}\mathbf{V}_{(2)}^\top\|_F^2.\end{aligned}$$

6.4 Proof of Lemma 5.7

$$\|\mathbf{U}_1 - \mathbf{U}_2 \mathbf{Q}\|_F^2 = 2k - 2\text{tr}(\mathbf{U}_1^\top \mathbf{U}_2 \mathbf{Q})$$

Let $\mathbf{U}_1^\top \mathbf{U}_2 = \mathbf{U} \mathbf{D} \mathbf{V}^\top$ be the singular value decomposition. Then $\mathbf{V} \mathbf{U}^\top \in O(k, k)$ and

$$\begin{aligned} \inf_{\mathbf{Q} \in O(k, k)} \|\mathbf{U}_1 - \mathbf{U}_2 \mathbf{Q}\|_F^2 &\leq 2k - 2\text{tr}(\mathbf{U}_1^\top \mathbf{U}_2 \mathbf{V} \mathbf{U}^\top) \\ &= 2k - 2\text{tr}(\mathbf{U} \mathbf{D} \mathbf{U}^\top) \\ &= 2k - 2\text{tr}(\mathbf{D}). \end{aligned}$$

On the other hand,

$$\begin{aligned} \|\mathbf{P}_{\mathbf{U}_1} - \mathbf{P}_{\mathbf{U}_2}\|_F^2 &= \|\mathbf{U}_1 \mathbf{U}_1^\top - \mathbf{U}_2 \mathbf{U}_2^\top\|_F^2 \\ &= 2k - 2\text{tr}(\mathbf{U}_1 \mathbf{U}_1^\top \mathbf{U}_2 \mathbf{U}_2^\top) \\ &= 2k - 2\text{tr}(\mathbf{U}_1^\top \mathbf{U}_2 \mathbf{U}_2^\top \mathbf{U}_1) \\ &= 2k - 2\text{tr}(\mathbf{D}^2). \end{aligned}$$

Since $\mathbf{U}_1, \mathbf{U}_2 \in O(p, k)$, $\|\mathbf{U}_1^\top \mathbf{U}_2\| \leq 1$ and therefore all the diagonal elements of \mathbf{D} is less than 1, which implies that $\text{tr}(\mathbf{D}) \geq \text{tr}(\mathbf{D}^2)$ and

$$\inf_{\mathbf{Q} \in O(k, k)} \|\mathbf{U}_1 - \mathbf{U}_2 \mathbf{Q}\|_F^2 \leq \|\mathbf{P}_{\mathbf{U}_1} - \mathbf{P}_{\mathbf{U}_2}\|_F^2.$$

6.5 Proof of Lemma 5.11

Without loss of generality, we can assume $\|\boldsymbol{\Sigma}\| = 1$ or else we can scale \mathbf{A} by $1/\sqrt{\|\boldsymbol{\Sigma}\|}$. Let $\mathbf{J} = \|\frac{1}{n} \mathbf{A}^\top \mathbf{A} - \boldsymbol{\Sigma}\|$ and by Lemma 5.10, there exists positive constants c_1, C_1 such that

$$P(\mathbf{J} \geq \max\{\delta, \delta^2\}) \leq e^{-c_1 t^2}, \quad \delta = C_1 \sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}.$$

Notice that $\mathbf{J} \geq 0$, then

$$\begin{aligned} \mathbb{E}[\mathbf{J}^4] &= \int_0^{+\infty} P(\mathbf{J} \geq x^{1/4}) dx \\ &= \int_0^{C_1^2 p^2/n^2} P(\mathbf{J} \geq x^{1/4}) dx + \int_{C_1^2 p^2/n^2}^1 P(\mathbf{J} \geq x^{1/4}) dx + \int_1^{+\infty} P(\mathbf{J} \geq x^{1/4}) dx \\ &\leq C_1^2 p^2/n^2 + \int_{C_1^2 p^2/n^2}^1 e^{-(\sqrt{n} x^{1/4} - C_1 \sqrt{p})^2} dx + \int_1^{+\infty} e^{-(\sqrt{n} x^{1/8} - C_1 \sqrt{p})^2} dx \\ &= C_1^2 p^2/n^2 + \int_{C_1^2 p^2/n^2}^1 4e^{-y^2} \left(\frac{y + C_1 \sqrt{p}}{\sqrt{n}} \right)^3 \frac{1}{\sqrt{n}} dy + \int_1^{+\infty} 8e^{-y^2} \left(\frac{y + C_1 \sqrt{p}}{\sqrt{n}} \right)^7 \frac{1}{\sqrt{n}} dy. \end{aligned}$$

There exists a large constant C_2 such that

$$\begin{aligned} \mathbb{E}[\mathbf{J}^4] &\leq C_1^2 p^2/n^2 + \frac{4}{n^2} \int_{C_1^2 p^2/n^2}^1 4e^{-y^2} C_2 (y^3 + p^{2/3}) dy + \frac{8}{n^4} \int_1^{+\infty} C_2 e^{-y^2} (y^7 + p^{7/2}) dy \\ &\leq C_1^2 p^2/n^2 + \frac{4}{n^2} \int_0^{+\infty} 4e^{-y^2} C_2 (y^3 + p^{2/3}) dy + \frac{8}{n^4} \int_0^{+\infty} C_2 e^{-y^2} (y^7 + p^{7/2}) dy. \end{aligned}$$

Notice that $\int_0^{+\infty} e^{-y^2} y^k dy$ is bounded for any $k \in \mathbb{Z}_+$ and $n \geq p$. There exists a large constant C_3 such that

$$\mathbb{E}[\mathbf{J}^4] \leq C_3 \frac{p^2}{n^2}.$$

References

- Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics* 34(1), 122–148.
- Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (second ed.). New York, NY: Wiley.
- Anderson, T. W. (1999). Asymptotic theory for canonical correlation analysis. *Journal of Multivariate Analysis* 70(1), 1–29.
- Arora, R. and K. Livescu (2013). Multi-view cca-based acoustic features for phonetic recognition across speakers and domains. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7135–7139. IEEE.
- Cai, T., Z. Ma, and Y. Wu (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability theory and related fields* 161(3-4), 781–815.
- Cai, T. T., Z. Ma, Y. Wu, et al. (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics* 41(6), 3074–3110.
- Cai, T. T. and A. Zhang (2016). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *arXiv preprint arXiv:1605.00353*.
- Chaudhuri, K., S. M. Kakade, K. Livescu, and K. Sridharan (2009). Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pp. 129–136. ACM.
- Chen, M., C. Gao, Z. Ren, and H. H. Zhou (2013). Sparse cca via precision adjusted iterative thresholding. *arXiv preprint arXiv:1311.6186*.
- Chen, X., H. Liu, and J. G. Carbonell (2012). Structured sparse canonical correlation analysis. In *International Conference on Artificial Intelligence and Statistics*, pp. 199–207.
- Dhillon, P. S., D. Foster, and L. Ungar (2011). Multi-view learning of word embeddings via cca. In *Advances in Neural Information Processing Systems (NIPS)*, Volume 24.
- Faruqui, M. and C. Dyer (2014). Improving vector space word representations using multilingual correlation. Association for Computational Linguistics.
- Fiedler, M. (2010). Notes on hilbert and cauchy matrices. *Linear Algebra and its Applications* 432(1), 351–356.

- Foster, D. P., R. Johnson, S. M. Kakade, and T. Zhang (2008). Multi-view dimensionality reduction via canonical correlation analysis. Technical report.
- Friman, O., M. Borga, P. Lundberg, and H. Knutsson (2003). Adaptive analysis of fmri data. *NeuroImage* 19(3), 837–845.
- Fukumizu, K., F. R. Bach, and M. I. Jordan (2009). Kernel dimension reduction in regression. *The Annals of Statistics*, 1871–1905.
- Gao, C., Z. Ma, Z. Ren, H. H. Zhou, et al. (2015). Minimax estimation in sparse canonical correlation analysis. *The Annals of Statistics* 43(5), 2168–2197.
- Gao, C., Z. Ma, and H. H. Zhou (2014). Sparse cca: Adaptive estimation and computational barriers. *arXiv preprint arXiv:1409.8565*.
- Gong, Y., Q. Ke, M. Isard, and S. Lazebnik (2014). A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision* 106(2), 210–233.
- Hom, R. A. and C. R. Johnson (1991). Topics in matrix analysis. *Cambridge UP, New York*.
- Hotelling, H. (1936). Relations between two sets of variables. *Biometrika* 28, 312–377.
- Hsu, P. (1941). On the limiting distribution of the canonical correlations. *Biometrika* 32(1), 38–45.
- Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis* 5(2), 248–264.
- Kakade, S. M. and D. P. Foster (2007). Multi-view regression via canonical correlation analysis. In *In Proc. of Conference on Learning Theory*.
- Kim, T.-K., S.-F. Wong, and R. Cipolla (2007). Tensor canonical correlation analysis for action classification. In *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*, pp. 1–8. IEEE.
- Loy, C. C., T. Xiang, and S. Gong (2009). Multi-camera activity correlation analysis. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1988–1995. IEEE.
- Ma, Z., Y. Lu, and D. Foster (2015). Finding linear structure in large datasets with scalable canonical correlation analysis. In *Proceedings of The 32nd International Conference on Machine Learning*, pp. 169–178.
- Mathias, R. (1993). The hadamard operator norm of a circulant and applications. *SIAM journal on matrix analysis and applications* 14(4), 1152–1167.
- Rasiwasia, N., J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos (2010). A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 251–260. ACM.

- Rudelson, M. and R. Vershynin (2013). Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.* 18, no. 82, 1–9.
- Sridharan, K. and S. M. Kakade (2008). An information theoretic framework for multi-view learning. In R. A. Servedio and T. Zhang (Eds.), *COLT*, pp. 403–414. Omnipress.
- Szarek, S. J. (1982). Nets of grassmann manifold and orthogonal group. In *Proceedings of research workshop on Banach space theory (Iowa City, Iowa, 1981)*, Volume 169, pp. 185.
- Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.
- Wang, W., R. Arora, K. Livescu, and J. Bilmes (2015). On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 1083–1092.
- Wedin, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics* 12(1), 99–111.
- Wedin, P. Å. (1983). On angles between subspaces of a finite dimensional inner product space. In *Matrix Pencils*, pp. 263–285. Springer.
- Witten, D. M., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, kxp008.
- Xu, C., D. Tao, and C. Xu (2013). A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.
- Yu, B. (1997). Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer.